# Alternative splicing and single-cell RNA-sequencing: a feasibility assessment

Jennifer Westoby

January 2020

Darwin College

This dissertation is submitted for the degree of Doctor of Philosophy.

# Preface

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit of 60,000 words.

# Summary

We know little about how isoform choice is regulated in individual cells for most spliced genes. In theory, single-cell RNA-sequencing (scRNA-seq) could enable us to investigate isoform choice at cellular resolution. Therefore, scRNA-seq could give insight into the fundamental molecular biology process of how alternative splicing is regulated within cells. However, scRNA-seq is a relatively new technology, and at the start of my PhD it was not clear whether existing bioinformatics approaches would enable accurate splicing analyses. In my PhD I consider what the limitations are when attempting to study alternative splicing using scRNA-seq and what can be done to overcome them.

Alternative splicing is commonly analysed using bulk RNA sequencing (bulk RNA-seq) data with isoform quantification software. It was not clear whether isoform quantification software designed for bulk RNA-seq would perform well when run on scRNA-seq data. To address this, I performed a simulation-based benchmark of isoform quantification software developed for bulk RNA-seq when run on scRNA-seq. I made two important findings. Firstly, I found that isoform quantification software performs poorly when run on Drop-seq data, but performs better when run on scRNA-seq data generated using full-length transcript protocols (eg. SMART-seq and SMART-seq2). Secondly, I found that for the most part, isoform quantification software performs almost as well when run on full-length scRNA-seq as it does when run on bulk RNA-seq. Based on these findings, I concluded that software tools to accurately quantify the reads from full-length scRNA-seq experiments exist, theoretically enabling alternative splicing to be analysed using scRNA-seq.

Encouraged by this result, I embarked on a series of experiments designed to

answer questions such as 'How many isoforms does a gene typically produce per cell?'. This is a key basic biology question that could in theory be answered using scRNA-seq. Unfortunately, I found that the results of these experiments were largely impossible to interpret because I was unable to distinguish between biological signal and technical noise. I realised that without a solid understanding of the technical noise and confounding factors associated with scRNA-seq, distinguishing biological signal from technical noise would be challenging and might not be possible. To address this, I embarked on a second simulation-based study, this time investigating the impact of technical noise on our ability to study alternative splicing using scRNA-seq. I simulated four situations: a situation where every gene expressed one isoform per cell, a situation where all genes expressed two isoforms per cell, a situation where all genes expressed three isoforms per cell and a situation where all genes expressed four isoforms per cell. Importantly, I explicitly simulated isoform choice, dropouts and quantification errors. The results of the four simulated situations were not trivial to distinguish from each other, raising concerns about the feasibility of resolving the more complex splicing patterns that probably exist in reality using scRNA-seq data. I concluded that attempts to study alternative splicing using scRNA-seq are currently substantially confounded by a high rate of dropouts and a lack of understanding about the mechanism of isoform choice. Importantly, improvements to isoform quantification software accuracy alone were insufficient to correct for confounding effects caused by dropouts. I propose that to enable accurate alternative splicing analyses using scRNA-seq, further research into accurately modelling dropouts is required, or alternatively, scRNA-seq technologies should be improved to increase their capture efficiency. Additionally, research into how isoform choice is regulated at a cellular level is necessary to enable accurate analyses. Overall, I find that it is not currently possible to accurately perform alternative splicing analyses using scRNA-seq. However, I am optimistic that with further research, it may become possible in the future.

# Acknowledgements

I am hugely grateful to my two supervisors, Anne Ferguson-Smith and Martin Hemberg, for their support and kindness over the years. Thank you both for seeing potential in me and for giving me the freedom to design my own research projects. I have learnt a huge amount over the last three years and my PhD would have been very different without the freedoms and opportunities you gave me. I would also like to thank all of the members of the Ferguson-Smith and Hemberg labs, who have taught me a huge amount throughout my PhD. We had a lot of fun together too! Thanks also to my PhD advisor, Steve Russell, for his support and pragmatic approach.

I would like to thank the BBSRC, and indirectly the UK taxpayer, for funding my PhD. Although I have been known to complain about the size of my stipend, I am hugely grateful that I have had the opportunity to be paid to learn and research. The UK taxpayer has always been at the back of my mind, and I hope that I have delivered good value for money through the research I have carried out. Needless to say, this PhD would not have been possible without the taxpayer's support.

Thank you to my parents, Marie and Richard, for your eternal support and unconditional love. You have always done your utmost to give me the best opportunities in life, and I am forever grateful. I would also like to thank my sister, Natalie, for the great times we have had together. I am hugely proud of Natalie's dedication to improving the lives of others in more direct ways than I will ever achieve. Mum, Dad and Natalie, you cannot choose your family, and so I feel incredibly lucky that you are mine.

My best friend, life partner and husband has supported me emotionally, intellec-

tually, fiscally, and in pretty much every other way imaginable throughout my PhD. Tom, I am forever grateful for your kindness, and so glad that you foolishly agreed to marry me. I am happiest each morning when I wake up and see you next to me.

I first became interested in learning to code at the age of 19. At this point, I was already enrolled in an undergraduate Natural Sciences course, and there were limited opportunities to incorporate programming into my studies. People often talk about the importance of women in tech having female mentors, however given around 10% of people in tech roles are female, the feasibility of this is questionable. Indeed, the three people who I credit most with enabling me to take my programming from an interest to a profession are all men. Thank you to Thomas Parks, Joseph Gardner and Martin Hemberg for your patience, kindness and wisdom. It is thanks to your support that I can proudly say that I have become a part of that 10%. Now that my feet are under the table, I am going nowhere.

# Contents

# 1

# Introduction

*I never am really satisfied that I understand anything; because, understand it well as I may, my comprehension can only be an infinitesimal fraction of all I want to understand...*

– Ada Lovelace, quoted by (Henderson, 1995).

## 1.1 Motivation

Alternative splicing is a fundamental process in molecular biology which enables multiple proteins to be generated from single genes. Alternative splicing is implicated in many biological processes, including the activation of pluripotency genes (Gabut et al., 2011), cell fate decisions (Yamazaki et al., 2018) and transcriptional regulation (Li et al., 2017). In addition, errors in alternative splicing have been implicated in numerous human diseases, including cancer (David and Manley, 2010) and genetic diseases such as Duchenne Muscular Dystrophy (Muntoni et al., 2003). Therefore, furthering our understanding of alternative splicing would be highly beneficial to our understanding of basic biology and human disease.

When bulk RNA-seq was first developed over a decade ago (Emrich et al., 2007; Lister et al., 2008), it enabled researchers to quantitatively measure transcript abundance on a genome wide scale. For the first time, researchers were able to study

alternative splicing across the entire transcriptome in a quantitative manner. However, although bulk RNA-seq enabled researchers to study more genes in a single experiment than had previously been possible using array based technologies, in one sense the resolution of bulk RNA-seq experiments remains poor. Consider Figure 1.1.

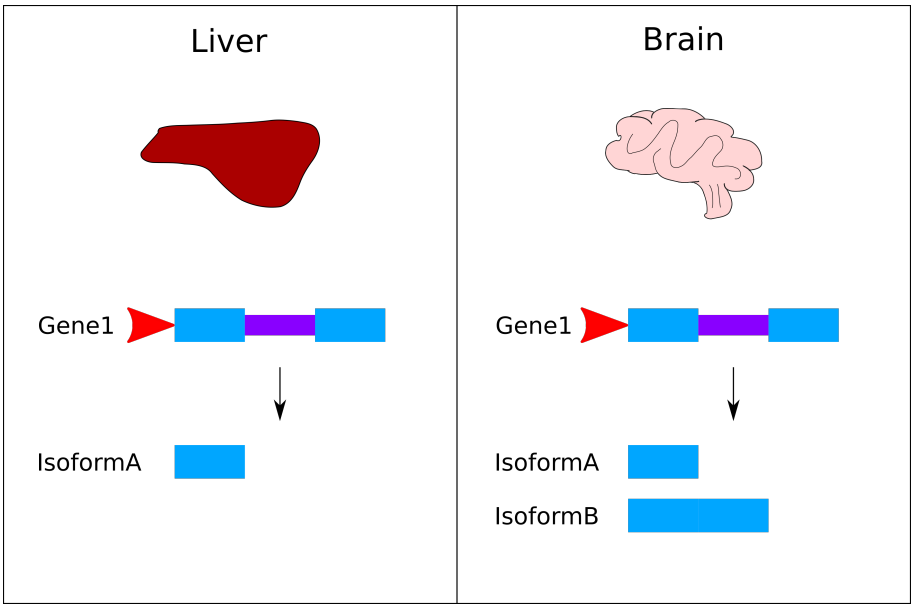


Figure 1.1: An imaginary bulk RNA-seq experiment in the liver and brain.

In Figure 1.1, we imagine an experiment in which we perform bulk RNA-seq on a mouse liver and brain. We find that for our gene of interest, 'Gene1', only one isoform ('Isoform1') is detected in the liver, whereas two isoforms ('Isoform1' and 'Isoform2') are detected in the brain. This is an informative result which indicates that there is differential splicing of Gene1 in the brain and the liver. However, because bulk RNA-seq is performed on a population of cells, based on this result there is very little that we can infer about how this differential splicing is regulated at a cellular level. For example, based on our bulk RNA-seq results, it would be challenging or impossible to answer the following questions:

1. Does Gene1 expresses two isoforms in every cell in the brain, or do some brain

cells exclusively express Isoform1 and others exclusively express Isoform2, or is there is some intermediate situation between these two extremes?

2. If there is heterogeneity in isoform expression between cells, does the heterogeneity correlate with cell type?

3. Do all cells express Gene1 in the brain or the liver?

At first sight, these questions appear deceptively simple. However, they are fundamental questions in molecular biology. If we are unable to answer them, we are severely limited in our understanding both of how alternative splicing is regulated in individual cells and of how splicing is regulated across complex tissues and organs.

In theory, single-cell RNA-seq (scRNA-seq) could enable us to answer all of the questions posed above, and more, by allowing us to study alternative splicing at a cellular resolution. That is, of course, if it is possible to study alternative splicing using scRNA-seq.

## 1.2 What is alternative splicing?

Alternative splicing is the process by which more than one mRNA transcript can be produced from a single gene. An example is illustrated in Figure 1.2.

Figure 1.2 illustrates a gene with a promoter, two exons and an intron. The promoter is the region where transcription begins, the exons are coding regions of the gene and the intron is non coding sequence. The gene in Figure 1.2 is transcribed and spliced to produce two mRNA isoforms, one consisting of only the first exon, and one consisting of the first and the second exon with the intron removed. The process by which two distinct isoforms can be produced from a single gene is known as alternative splicing.

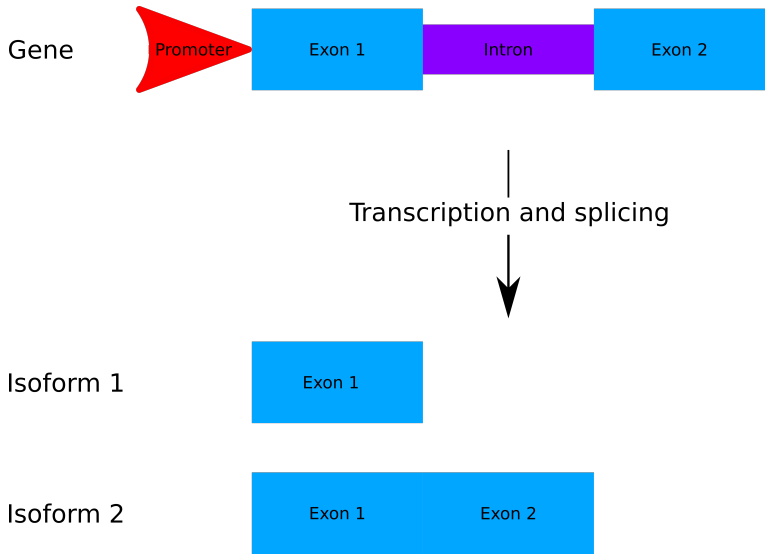


Figure 1.2: An example of an alternatively spliced gene.

## 1.3 Molecular mechanisms of alternative splicing

The previous section described the outcome of alternative splicing. In this section I will briefly explain how alternative splicing takes place at a molecular level.

Alternative splicing is catalysed by the spliceosome. The spliceosome is a ribonucleoprotein (RNP) complex made up of five spliceosomal RNP subunits and many more protein cofactors. snRNAs, the RNA molecules that form part of the spliceosomal subunits, are thought to provide the catalytic activity required for alternative splicing (Matera and Wang, 2014). This hypothesis is supported by experiments in which splicing-like reactions have been catalysed in protein free systems (Valadkhan et al., 2007, 2009).

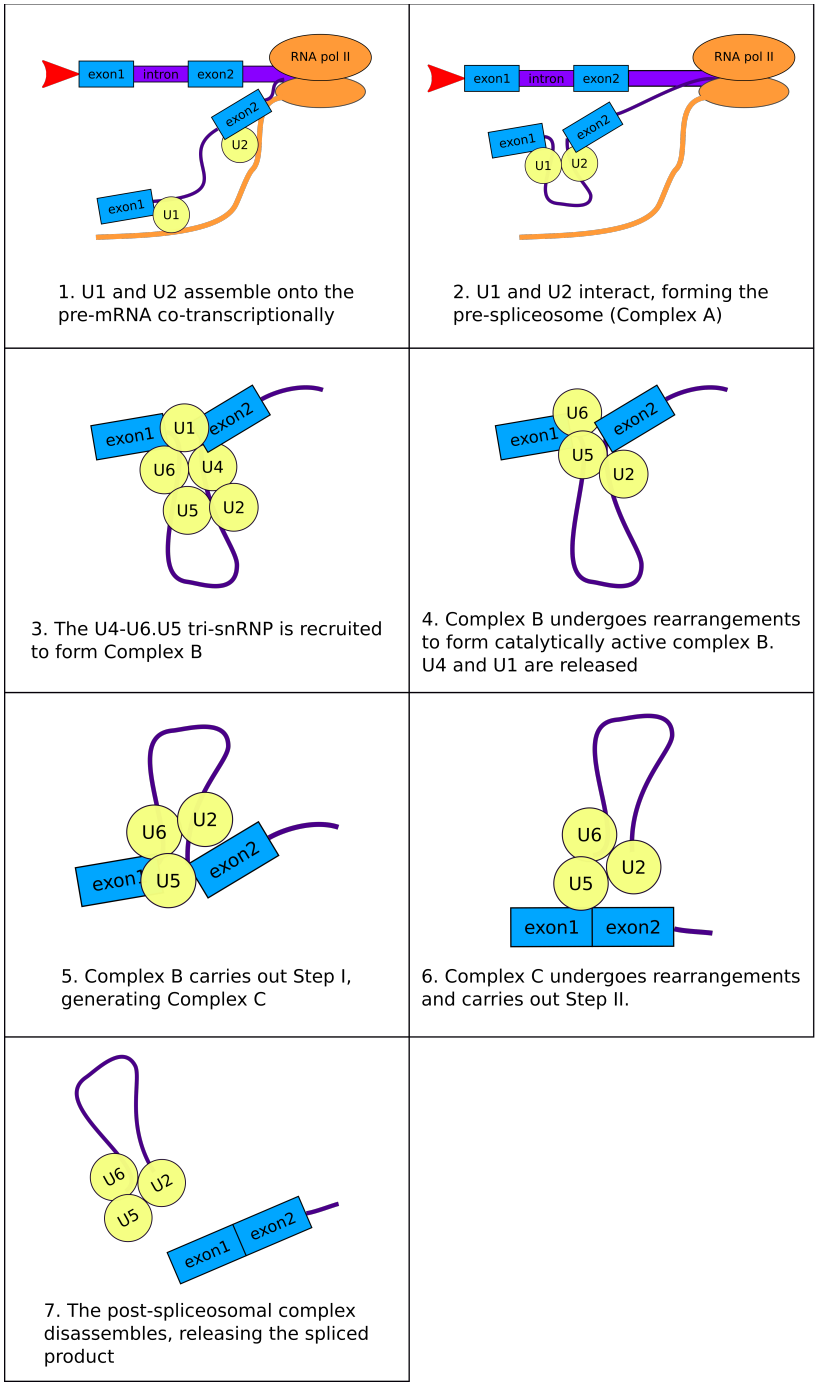


Figure 1.3: A schematic of the alternative splicing process. Adapted from Matera et al. (Matera and Wang, 2014)

Figure 1.3 illustrates the splicing process from start to end. The catalytic process of alternative splicing can be considered to take place in two steps, however prior to catalysis the spliceosome must assemble. This begins with two of the spliceosomal subunits, U1 and U2, assembling onto the pre-mRNA co-transcriptionally. The Carboxy-Terminal Domain (CTD) of RNA polymerase II possibly mediates this process (Görnemann et al., 2011; Wiesner et al., 2002; Morris and Greenleaf, 2000). Next, U1 and U2 interact to form the pre-spliceosome, also known as Complex A. The U4-U6.U5 tri-snRNP complex is then recruited to Complex A, leading to the formation of Complex B. Complex B releases two spliceosomal subunits, U1 and U4 (Raghunathan and Guthrie, 1998), and undergoes rearrangements to form catalytically active Complex B*. The spliceosome is now finally ready to begin catalysing a splicing reaction. The first step of the splicing reaction (Step I or 'branching') generates Complex C. Complex C contains free exon1 and the intron-exon2 lariat intermediate. Complex C then undergoes rearrangements and carries out Step II, leading to the formation of the post spliceosomal complex. The post spliceosomal complex contains the lariat intron and the spliced together exons. The post spliceosomal complex then dissembles, releasing the spliced product.

A question that is likely to arise from this description is how the spliceosome is able to recognise where it should assemble on the pre-mRNA and thus which sequences it should splice out and which sequences should be retained in the final mature mRNA. The DNA sequence at splice sites (the locations on the pre-mRNA where the splicesome assembles) is not random. In the nucleus, introns typically begin with a GT and end with an AG (Breathnach et al., 1978). As the number of splice sites studied increased, it was recognised that there were a number of consensus splice site sequences, although the consensus sequences are not universal and variations have been observed (Mount, 1982). With the advent of genomic sequencing, software to predict splice sites based on genomic sequence was developed. Whilst this has been extremely useful in terms of enabling scientists to study alternative splicing on a genome wide scale, predicting splice sites based on sequence is not trivial and is likely to be error prone (Mount, 2000). However, although predicting splice sites based on sequence is challenging for humans and computers, the sequence

at splice sites is recognised by components of the spliceosome and thus determines where alternative splicing should take place. The 5' splice site, located at the start of the intron, is recognised by the U1 snRNP, both by base pairing with the U1 snRNA and in a base pairing independent manner by the U1C subunit of U1 snRNP (Du and Rosbash, 2002). The 3' splice site, located at the end of the intron, is recognised by the U2 snRNP and other associated splicing factors (Matera and Wang, 2014).

## 1.4 History of alternative splicing

Alternative splicing was independently discovered in 1977 by Berget et al. and Chow et al. (Berget et al., 1977; Chow et al., 1977). The discovery was made by studying hybridisation reactions between single stranded adenovirus DNA and its complementary RNA. The formation of single stranded DNA loops, interspersed between lengths of RNA-DNA hybridisation, were observed by electron microscopy. These loops corresponded to introns. Based on these observations, a mechanism in which lengths of RNA were 'spliced' out of mature mRNAs was proposed to explain the observation that the length of DNA sequence corresponding to a gene is often much longer than the length of the transcribed mRNA (Berget et al., 1977).

Two years later, a study by Lerner et al. into two antibodies produced by systemic Lupus patients found that the antibodies pulled down six snRNPs (Lerner and Steitz, 1979). Lerner et al. identified that the six snRNPs formed a complex with one another, likely making this the first time that the spliceosome was experimentally isolated. The observation that one of the six snRNP's (U1's) RNA has a complementary nucleotide sequence to many splice sites led some to suggest that the isolated snRNP complex might play a role in alternative splicing (Lerner et al., 1980). In the years that followed, further studies confirmed that the other snRNPs in the complex played a role in splicing (Black et al., 1985; Krainer and Maniatis, 1985; Berget and Robberson, 1986; Grabowski and Sharp, 1986).

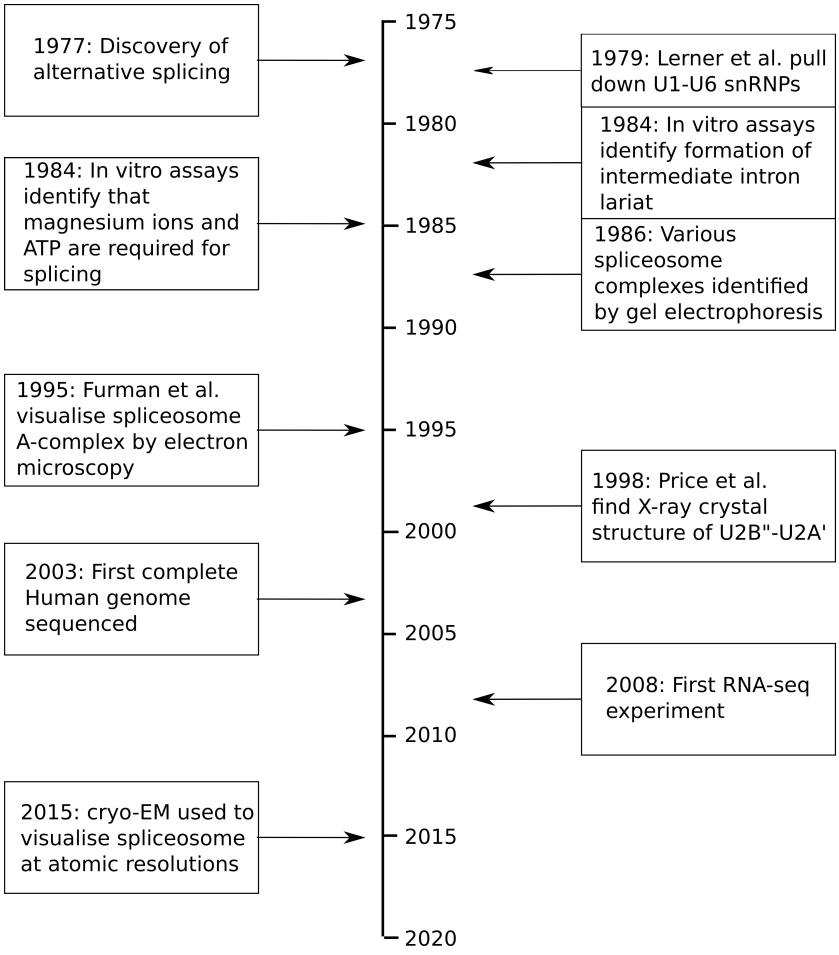


Figure 1.4: A timeline of the major events in splicing research.

Development of in vitro splicing assays enabled further biochemical insights into the process of alternative splicing (Shi, 2017). In vitro assays identified that ATP and Mg2+ were necessary for alternative splicing (Hernandez and Keller, 1983; Hardy et al., 1984) and confirmed the existence of an intron lariat intermediate (Grabowski et al., 1984; Padgett et al., 1984; Ruskin et al., 1984). Non-denaturing gel electrophoresis experiments identified that the spliceosome formed several distinct complexes throughout the splicing reaction (Lamond et al., 1987; Konarska and Sharp, 1986, 1987; Pikielny et al., 1986; Cheng and Abelson, 1987; Bindereif and Green,

1987).

Further biochemical experiments studying the interactions and spatial location of spliceosome components enabled researchers to build a simple 'map' of the spliceosome and furthered the field's mechanistic understanding of the chemical processes behind alternative splicing (Shi, 2017; Newman and Norman, 1991; Madhani and Guthrie, 1992; Wassarman and Steitz, 1992; Wyatt et al., 1992; Lesser and Guthrie, 1993; Sontheimer and Steitz, 1993; Anokhina et al., 2013; Newman et al., 1995). Microscopy based approaches were also used to study the structure of the spliceosome. Due to the highly dynamic nature of the spliceosome, it is almost impossible to crystallise spliceosome complexes (Shi, 2017). Consequently, X-ray crystallography based approaches have had some success in studying some of the subcomplexes and components of the spliceosome (Sickmier et al., 2006; Lin and Xu, 2012; Jenkins et al., 2013; Yoshida et al., 2015; Leung et al., 2011; Zhou et al., 2014; Montemayor et al., 2014; Weber et al., 2010; Pomeranz Krummel et al., 2009; Kondo et al., 2015; Price et al., 1998), but a fully assembled spliceosome has never been successfully crystalised. Electron microscopy approaches have had more success, and have been used to successfully study the structure of many of the spliceosomal complexes (Behzadnia et al., 2007; Furman and Glitz, 1995; Boehringer et al., 2004; Wolf et al., 2009; Deckert et al., 2006; Bessonov et al., 2010; Golas et al., 2010; Jurica et al., 2004; Ilagan et al., 2013; Fabrizio et al., 2009; Ohi et al., 2007; Chen et al., 2014). However, these studies only had moderate resolutions, meaning the structural insight that could be gained from them was limited. A real breakthrough came in 2015, when cryo-EM studies of the spliceosome enabled structural study of the spliceosome at atomic level resolutions (Yan et al., 2015; Hang et al., 2015). The cryo-EM studies of 2015 and after have given researchers a new level of mechanistic insight into the biochemical splicing reaction (Yan et al., 2015; Hang et al., 2015; Agafonov et al., 2016; Yan et al., 2016; Wan et al., 2016b,a; Galej et al., 2016; Yan et al., 2017; Fica et al., 2017; Rauhut et al., 2016; Bertram et al., 2017; Nguyen et al., 2016; Zhang et al., 2018; Haselbach et al., 2018; Zhang et al., 2019).

In parallel to the biochemical study of the mechanistic process of alternative splicing, since the advent of whole genome sequencing alternative splicing has also

been studied at a genomic level. Initially, DNA sequence was annotated using gene annotation software, which made predictions about gene structure including splice sites and promoters (Mount, 2000; Reese et al., 2000). Later, reads from RNA-seq experiments enabled researchers to verify gene and isoform structure predictions, make new predictions, and to quantify the relative abundance of each splice isoform (Weber, 2015). Whilst methods for quantifying isoform abundance or predicting gene structure based on sequencing reads remain far from perfect (Hölzer and Marz, 2019; Everaert et al., 2017), it is fair to say that sequencing technologies have given us a level of insight into splicing which was previously unimaginable.

## 1.5 Alternative splicing and disease

Alternative splicing defects have been linked to many human diseases, including Duchenne Muscular Dystrophy (Muntoni et al., 2003), Early Onset Parkinson Disease (Samaranch et al., 2010), Retinitis pigmentosa (Tanackovic et al., 2011; Cvačková et al., 2014) and cancer (David and Manley, 2010). Furthering our understanding of how splicing is regulated at a cellular level is likely to be relevant to the therapeutic treatment of many diseases in which splicing errors play a role.

It is perhaps unsurprising that splicing has been implicated in such a wide range of diseases given that a study by Lopez-Bigas et al. suggests that in humans, up to 60% of disease causing mutations affect splicing (López-Bigas et al., 2005). Related studies suggested that one third of all mutations (Lim et al., 2011), and one quarter of all coding mutations (Sterne-Weiler et al., 2011) impact splicing. Clearly these numbers can differ quite substantially depending on the type of mutation investigated and the methodology used, however they do all support the hypothesis that a high proportion of mutations alter splicing. An additional hypothesis for why splicing is implicated in so many human diseases is that mutations which alter splice sites can dramatically change the protein produced from the parent gene. In contrast, non-splice site mutations have a moderate likelihood of having a small or no impact on the protein structure produced. Such mutations are unlikely to produce a disease phenotype.

## 1.6 What is scRNA-seq?

scRNA-seq is a relatively new sequencing technology in which RNA molecules are captured from individual cells and sequenced. scRNA-seq has the potential to give deep insight into how alternative splicing is regulated in individual cells across the entire transcriptome. However, this is only possible if splicing analyses are not confounded by the high degree of technical noise in scRNA-seq data. I will begin my discussion of the current state of scRNA-seq based approaches by considering the history of scRNA-seq.

## 1.7 History of scRNA-seq

Figure 1.5 is a timeline of some of the major events in the scRNA-seq's short history. Our timeline begins with the first published scRNA-seq experiment in 2009, when Tang et al. performed scRNA-seq on a handful of mouse blastomeres and oocytes (Tang et al., 2009). In their study, Tang et al. noted that 8-19% of known genes with two or more isoforms expressed at least two isoforms in individual cells, showing that there has been an interest in studying alternative splicing in individual cells from the first scRNA-seq experiments.

Our next major event comes a year later, when Guo et al. took an array based approach to profile the expression of 48 genes in approximately 500 cells (Guo et al., 2010). Although this was not an scRNA-seq experiment, it had an important impact on the single cell community. Guo et al. demonstrated that they could use the expression of their 48 profiled genes to identify each cell's cell type. The idea that a cell's identity can be determined by the expression of a selection of its genes is a key part of the philosophy of modern scRNA-seq clustering algorithms.

In 2011, Islam et al. developed STRT-seq, a multiplexed scRNA-seq protocol (Islam et al., 2011). This was revolutionary because it enabled researchers to sequence hundreds or thousands of cells in a single experiment, whereas previously most scRNA-seq experiments sequenced tens of cells, if not fewer (Tang et al., 2009, 2010). A year later, an scRNA-seq protocol called SMART-seq was published (Ramsköld

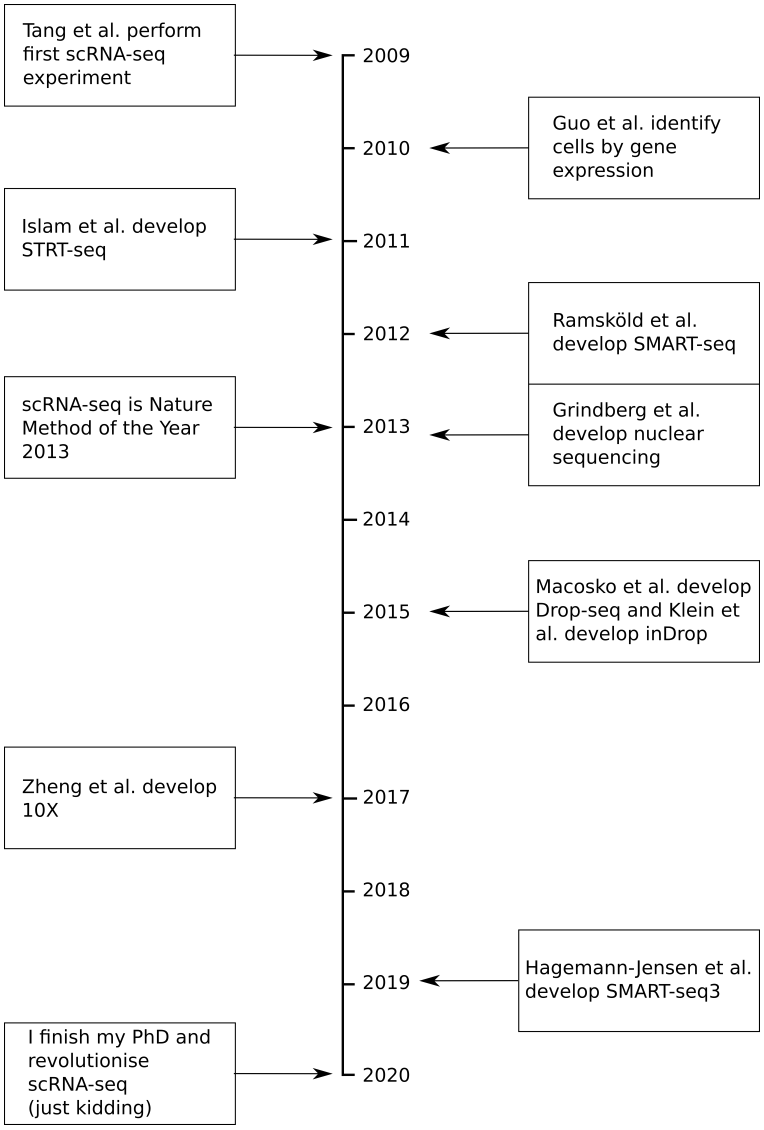


Figure 1.5: Timeline of selected events in the history of scRNA-seq.

et al., 2012). SMART-seq was the first 'full-length' scRNA-seq protocol, so called because the protocol attempted to sample reads across the full length of transcripts and thus reduce the 3' bias observed in previous protocols. In practice, SMART-seq is not entirely free from 3' bias although it is reduced (Picelli, 2017). A newer protocol called SMART-seq2 has further reduced coverage bias, although again the bias

is not fully eliminated (Picelli et al., 2014). The development of 'full-length' protocols is important for studying alternative splicing using scRNA-seq. Protocols which only sequence one end of each transcript are unable to distinguish between isoforms whose exon usage differs only at the other end of the transcript. Such protocols are therefore of limited use for alternative splicing studies.

It is challenging to obtain intact isolated cells from some tissues, such as neural progenitor cells and dental gyrate. Consequently, scRNA-seq can be difficult in these tissues. This problem was overcome in 2013 by Grindberg et al., who developed an scRNA-seq protocol in which nuclei rather than whole cells were sequenced (Grindberg et al., 2013). This enabled neurobiologists to apply scRNA-seq to their studies in a widespread manner. It was also exciting news for those interested in studying alternative splicing, as alternative splicing is known to be widespread in the human brain and to correlate with brain development and neuronal differentiation (Su et al., 2018).

2013 could be regarded as a turning point in scRNA-seq's history. Following technological advances in 2013 and in previous years, scRNA-seq was crowned Nature Method of The Year 2013 (NatMethods, 2014). By the start of 2014, scRNA-seq had advanced to a point where a sufficiently large number of cells could be sequenced to perform meaningful statistical analyses (ie. over one hundred cells) and the experimental protocols were reasonably accessible for many labs. This heralded the beginning of an explosion in the number of scRNA-seq publications.

The next major event on our timeline is the development of droplet based scRNA-seq methods. Droplet based scRNA-seq methods capture cells for sequencing in droplets and changed the field by enabling hundreds of thousands and eventually millions of cells to be sequenced in a single experiment. The first two droplet based methods were Drop-seq and InDrop, which were published in the same issue of Cell in 2015 (Klein et al., 2015; Macosko et al., 2015). Two years later, 10X Genomics released their droplet based protocol (Zheng et al., 2017). Due to the low cost and user friendly nature of 10X's platform, 10X is currently the dominant player in scRNA-seq.

Whilst this thesis was being written, a new scRNA-seq library preparation proto-

col called SMART-seq3 was released (Hagemann-Jensen et al., 2019). SMART-seq3 is the first technology to combine full length reads and reads containing Unique Molecular Identifiers (UMIs) and has therefore been greeted with excitement by the scRNA-seq community.

At the time of writing, a range of scRNA-seq protocols are widely in use and new protocols are still being developed. Identifying which of these protocols are likely to be applicable to studying alternative splicing is an important goal of my thesis.

## 1.8 A generalised scRNA-seq protocol

Figure 1.6 is a flowchart for a generalised scRNA-seq protocol. In this section I will discuss each step of the generalised protocol and how some steps differ between protocols.

The first step of our generalised protocol is cell capture and lysis. Broadly speaking, scRNA-seq protocols can be split into three groups depending on their mechanism of cell capture:

1. Microwell based capture methods. In these methods, cells are captured in wells using lasers, pipettes or Fluorescent Activated Cell Sorting (FACS) (Lafzi et al., 2018; Svensson et al., 2018). Using FACS for sorting has two major advantages - firstly FACS methods often lend themselves well to automation and secondly, FACS based methods allow researchers to enrich for cells of a particular cell type.

2. Microfluidics based capture methods. These technologies capture cells in nanoliter reaction volumes in an automated fashion, lending themselves well to medium sized experiments (order of 1000 cells) (Lafzi et al., 2018; Svensson et al., 2018).

3. Droplet based capture methods. These are the newest methods in which cells are captured in nanoliter droplet emulsions (Svensson et al., 2018). Very large

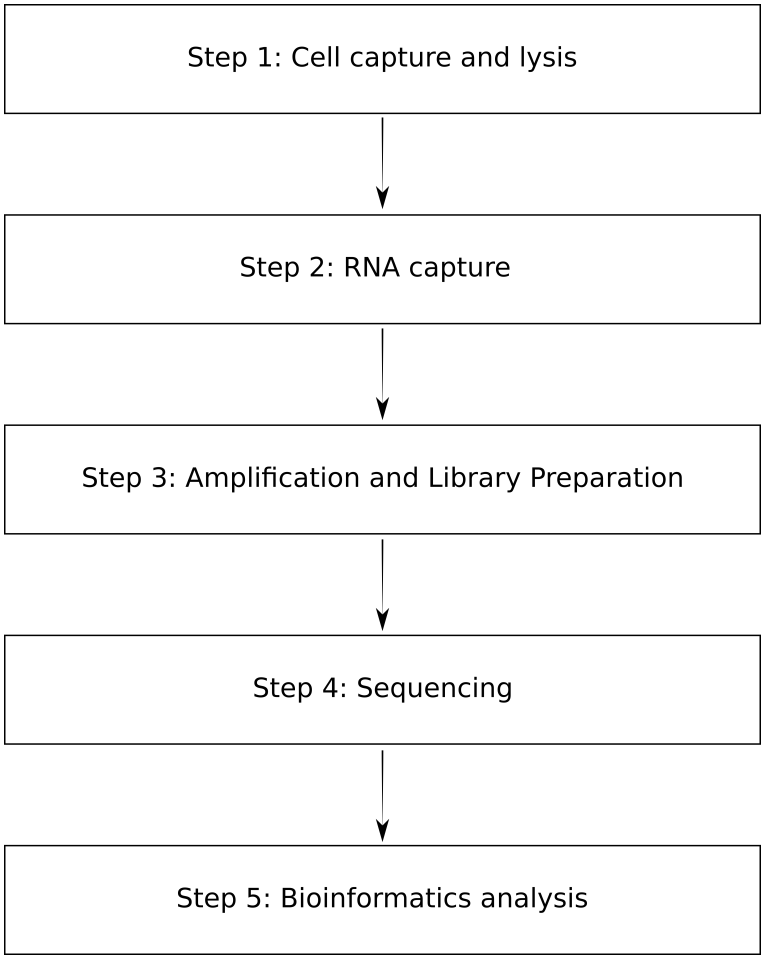


Figure 1.6: Flowchart of a generalised scRNA-seq protocol.

numbers of cells can be captured in these experiments (order of 10,000 cells and greater), making them best suited to large scale experiments.

Once they have been captured, cells are lysed in their wells/reaction volumes/droplets. In the next step of our protocol RNA is captured from each cell. All of the scRNA-seq protocols I will consider in my thesis capture poly(A) tailed mRNAs, typically using poly(T) oligonucleotides (Lafzi et al., 2018). The captured RNA is converted to cDNA prior to amplification in Step 3. All of the protocols considered in my thesis use PCR to amplify the harvested cDNA although protocols which use in vitro

transcription as their amplification method also exist (Svensson et al., 2018). Any final library preparation steps take place and the libraries are sequenced in Step 4. In Step 5, the obtained sequences can be computationally analysed. Typically this will involve some form of reads alignment and gene or isoform quantification, followed by a single cell specific analysis. Some typical analyses are considered in the next section.

## 1.9 Typical applications of scRNA-seq

In this section, I give a brief overview of a selection of popular scRNA-seq applications. All of the applications below are typically run using gene rather than isoform level expression estimates. In theory, these applications could be used with isoform level expression estimates. Uncertainty over the feasibility of isoform quantification in scRNA-seq and ability to detect isoforms in scRNA-seq data are two likely reasons that these applications are usually run with gene level expression estimates. In addition, gene level quantification is sufficient for many scRNA-seq experiments. For example, if the goal of an scRNA-seq experiment is to identify cell types in a blood sample, gene level quantification estimates are likely to enable accurate cell type identification.

### 1.9.1 Clustering

Guo et al.'s pioneering experiments identifying cells based on their gene expression gave rise to perhaps the most popular application of scRNA-seq (Guo et al., 2010). Clustering is now considered to be an essential step of many scRNA-seq experiments and is typically used to attempt to infer cell identities based on gene expression (Luecken and Theis, 2019; Petegrosso et al., 2019). This is not trivial. Technical noise is a major issue in scRNA-seq, and if not appropriately corrected for, technical noise rather than biological signal can dominate clusters (Luecken and Theis, 2019).

Clustering approaches and ideas about cell types have played a role in motivating large scale scRNA-seq projects such as the Human Cell Atlas. The goal of the Human

Cell Atlas project is to provide a catalogue of the cell types present in the human body and a map of the relationships between them (Regev et al., 2017). Projects such as the Human Cell Atlas are likely to lead to an increased understanding of what a typical human looks like at a cellular level, and may help answer questions such as: 'How many cell types are there in a typical human?'. In addition, the Human Cell Atlas will provide a valuable resource to the scientific community.

### 1.9.2 Pseudotime

It may not always be possible to separate a population of cells into discrete types. For example, a differentiating population of cells may lie along a continuum between two cell types, rather every cell being entirely one cell type or the other. Unlike clustering algorithms, which attempt to cluster cells into discrete cell types, pseudotime algorithms attempt to order cells along a linear or branching topological continuum based on similarity of expression (Saelens et al., 2019).

### 1.9.3 Differential expression

Identifying differences in gene expression between two populations of cells is a popular application of scRNA-seq. Differential expression analyses pre-existed scRNA-seq and a number of software tools for differential expression in bulk RNA-seq exist (Costa-Silva et al., 2017). However, there are substantial differences between bulk and scRNA-seq, most notably the much higher degree of technical noise present in scRNA-seq relative to bulk. Therefore it was not clear whether differential expression software designed for bulk RNA-seq would give accurate results when run on scRNA-seq (Luecken and Theis, 2019; Soneson and Robinson, 2018b). A recent benchmark found that bulk RNA-seq differential expression software did not perform more poorly on scRNA-seq than differential expression software that had been designed for scRNA-seq (Soneson and Robinson, 2018b). However, some bulk RNA-seq differential expression softwares were more sensitive to pre-filtering of genes than scRNA-seq softwares (Soneson and Robinson, 2018b).

### 1.9.4 Network modelling

Network modelling approaches represent biological systems as nodes and edges, where nodes can represent genes, proteins or other biological molecules, and edges can represent binding, enzymatic reactions, or other types of interaction (Blencowe et al., 2019). Network modelling is a classic systems biology approach that attempts to understand a complex system. Traditionally, network modelling has relied on bulk genomic data, and attempts to apply network modelling approaches to scRNA-seq data are relatively recent. A benchmark study found that network methods designed for bulk RNA-seq performed extremely poorly when run on scRNA-seq data (Chen and Mar, 2018). This is likely to reflect the substantial differences between bulk and scRNA-seq data. The high rate of technical dropouts, more dramatic batch effects and much larger dimensionality of scRNA-seq compared with bulk RNA-seq are possible reasons that bulk methods performed poorly on scRNA-seq (Blencowe et al., 2019). Despite these challenges, a large number of scRNA-seq methods for network analysis have been developed. However, in the benchmark above three scRNA-seq network analyses were considered in addition to the bulk approaches, and the performance of the scRNA-seq approaches was also generally poor (Chen and Mar, 2018). This indicates a need for further, larger benchmarks to establish whether scRNA-seq network methods with good performance exist. If they do not, research into which features of scRNA-seq data make network analyses so challenging would be informative.

## 1.10 Technical noise in scRNA-seq

A major issue when analysing scRNA-seq data is the high degree of technical noise relative to bulk RNA-seq. In this section I will discuss some of the main types of technical noise that are commonly seen in scRNA-seq. An understanding of the technical noise present in scRNA-seq data is crucial when analysing scRNA-seq data, as failing to recognise and correct for technical noise can lead to inappropriate analyses being performed, generating misleading results.

### 1.10.1 Read quality

As in any sequencing technology, issues relating to read quality are a source of technical noise in scRNA-seq. Tools such as FastQC can be used to investigate a variety of read quality metrics and establish the degree to which read quality is likely to be a significant source of technical noise (Andrews, 2015). Tools such as cutadapt can be used to trim unwanted sequencing adapters from reads, improving the accuracy of alignment (Martin, 2011).

### 1.10.2 Multiplets and empty wells

The goal of scRNA-seq is to sequence individual cells. However, the cell capture and isolation process is not perfect for any protocol. A common problem is that cells will sometimes clump together, so that multiple cells ('multiplets') end up in one well or droplet, depending on the sequencing protocol. Another common problem is that some wells or droplets contain no cells. These empty wells or droplets do still typically contain RNA, originating cells that lysed prematurely before capture.

As the goal of scRNA-seq is to sequence individual cells, multiplets and empty wells or droplets should be removed prior to analysing the data, leaving only 'singlets' (wells or droplets containing only one cell). There are a variety of methods for removing multiplets and empty wells or droplets from scRNA-seq datasets (Zappia et al., 2018). Perhaps the simplest method is to plot the dataset's distribution of library size and use the distribution to determine which cells are singlets and should be kept. If the dataset contains a large number of multiplets and empty cells, typically multiple peaks can be seen in the library size distribution. The first peak lies close to zero and corresponds to wells or droplets which didn't capture any cells. These empty cells should be removed. The second peak corresponds to the average library size of a singlet. These cells should be kept. Typically, the multiplet peaks will appear at multiples of the average library size of singlet. These multiplets should be removed.

More sophisticated tools for identifying multiplets and empty wells and droplets also exist (Ilicic et al., 2016; DePasquale et al., 2018; Wolock et al., 2019). If the

dataset under consideration contains genetically heterogeneous cells, genetic information can be used to resolve between singlets and multiplets (Kang et al., 2018). A new technology called Cell Hashing, in which oligo-tagged antibodies are used to label distinct cell populations, enables oligo information to be used to facilitate multiplet identification (Stoeckius et al., 2018).

### 1.10.3 Unwanted Biological Noise

Usually the aim of a scRNA-seq experiment is to detect biological signal. However, sometimes some of the biological noise present can confound the biological signal we are trying to detect. For example, during cell isolation and lysis, some cells become stressed and die. The stressed cells activate apoptosis pathways, altering their transcriptional profile. This is a strong biological signal with the potential to confound downstream analyses. Therefore, these cells should normally be removed. When dying cells lyse, cytoplasmic RNA leaks out of the cells, whereas mitochondrial RNA is relatively protected because it is encapsulated by the mitochondrial memberanes. Therefore, stressed and dying cells are commonly identified based on the proportion of each cell's sequencing library made up of mitochondrial reads (Luecken and Theis, 2019).

If an scRNA-seq dataset is made up of cells in a variety of cell cycle stages, it is common for clustering algorithms to cluster cells by their cell cycle stage, potentially hiding undetected subpopulations of cells. Buettner et al developed software that attempts to correct gene expression based on cell cycle stage, thus removing this confounder (Buettner et al., 2015).

### 1.10.4 Dropouts

Dropouts are a phenomenon where genes or isoforms for which high expression is detected in some cells are not detected as expressed at all in other cells (Kharchenko et al., 2014). This can occur for biological reasons. For example, if a mixture of cell types is sequenced, and a gene of interest is expressed in some cell types but not others, there will be biological dropouts in the scRNA-seq data. Alternatively, if a

gene has 'bursty' expression, so each cell sometimes switches that gene's transcription 'ON' and at other times does not express the gene at all, it is likely that the gene will exhibit biological dropouts in scRNA-seq data.

However, it is now known that a substantial proportion of dropouts in a typical scRNA-seq experiment are technical in origin. In this thesis, I will define the capture efficiency of an scRNA-seq experiment as the proportion of expressed genes (or isoforms) in a cell that are detected as expressed by scRNA-seq, ie.:

$$CaptureEfficiency = \frac{NumberOfGenesDetected}{NumberOfGenesExpressed}$$

It has been shown that the capture efficiency of scRNA-seq can be 10% or less (Marinov et al., 2014; Svensson et al., 2017; Islam et al., 2014). In an scRNA-seq experiment with a capture efficiency of 10%, only about 10% of transcripts in a cell generate sequencing reads. Consequently, a high percentage of expressed genes and isoforms will not be detected as expressed. Dropouts that occur due to a failure to capture reads from expressed transcripts are known as technical dropouts, and are a major source of technical noise. In addition to the low capture efficiency of scRNA-seq, technical dropouts can also occur if cells are very shallowly sequenced. Sequencing cells at low read depths is more common in droplet based protocols, and sometimes occurs due to financial considerations.

How best to correct for technical dropouts is an area of ongoing research. A common approach for attempting to correct for technical dropouts is imputation. Imputation approaches use mathematical modelling approaches to attempt to predict which dropouts are technical, and convert these zero values to their predicted 'true' values. A recent benchmark found that many imputation methods introduce false positives, illustrating that the problem of correcting for technical dropouts is not yet fully solved (Andrews and Hemberg, 2018b).

### 1.10.5 Batch effects

Batch effects occur due to technical variability between samples. Batch effects are not exclusive to scRNA-seq, they are also commonly corrected for in bulk RNA-seq (Fei et al., 2018). However batch effects are more complex and thus more challenging to correct for in scRNA-seq, where each cell can be regarded as a sample and therefore a batch (Tung et al., 2017).

Methods to correct for batch effects in scRNA-seq experiments exist, however they are only effective if the experiment has been appropriately designed. Importantly, if the different biological conditions being tested entirely overlap with batch effects, the experiment is confounded and it will be impossible to correct for batch effects. For example, if an experiment with two conditions, A and B, is performed, and cells in condition A are sequenced in one lab and cells in condition B are sequenced in another lab, it is impossible to determine whether differences between A and B are due to the condition or the lab. The experiment is therefore completely confounded by batch effects. If cells in each condition had been split between the two labs, batch correction would have been possible.

Reasons batch effects can occur in scRNA-seq include:

1. Samples were prepared in different facilities
2. Samples were prepared at different times
3. Samples were prepared by different people
4. Samples were sequenced at different depths
5. Samples were prepared using different reagents
6. Samples were sequenced in different lanes
7. Samples were prepared on different plates
8. Samples were prepared in different wells
9. Pipetting errors during library preparation

Reasons 1-7 can sometimes (though not always) be avoided with good experimental design. Reasons 8 and 9 can lead to batch effects between cells sequenced in the same run and are almost impossible to fully avoid.

Some of the reasons above generate more technical noise than others and are thus more important to correct for. For example, plate based and well based effects tend to be very minor, whereas having different people prepare samples can lead to more substantial noise, especially if their interpretation of how to follow the experimental protocol differs. Unfortunately, it is usually impossible to design a perfect experiment entirely unconfounded by batch effects, especially when dealing with rare or perishable samples. The goal should therefore be to minimise confounding factors as far as possible, and to recognise what confounders are likely to remain when interpreting results. A number of batch correction methods exist and should be used to correct for batch effects where possible. Based on a benchmark comparing scRNA-seq pipelines, Tian et al. recommended MNNs as a good general method (Tian et al., 2019; Haghverdi et al., 2018).

### 1.10.6 PCR amplification bias

Like batch effects, PCR amplification bias is not unique to scRNA-seq. PCR amplification bias exists in any protocol which involves a PCR amplification step. However, due to the small amount of starting material in scRNA-seq experiments, PCR amplification bias is more dramatic in scRNA-seq compared to eg. bulk RNA-seq (Smith et al., 2017).

Unique Molecular Identifiers (UMIs) have largely corrected for PCR amplification bias in some scRNA-seq protocols. UMIs are unique barcodes that are added to cDNA molecules prior to PCR amplification (see Figure 1.7) (Islam et al., 2014). After sequencing, the combination of the transcript identity and the UMI can be used to estimate how many copies of the transcript were originally captured.

Errors can occur during UMI deduplication, especially if there is a sequencing error in the UMI barcode or if the same UMI sequence binds to more than one cDNA molecule originating from the same gene. However, provided that the UMIs

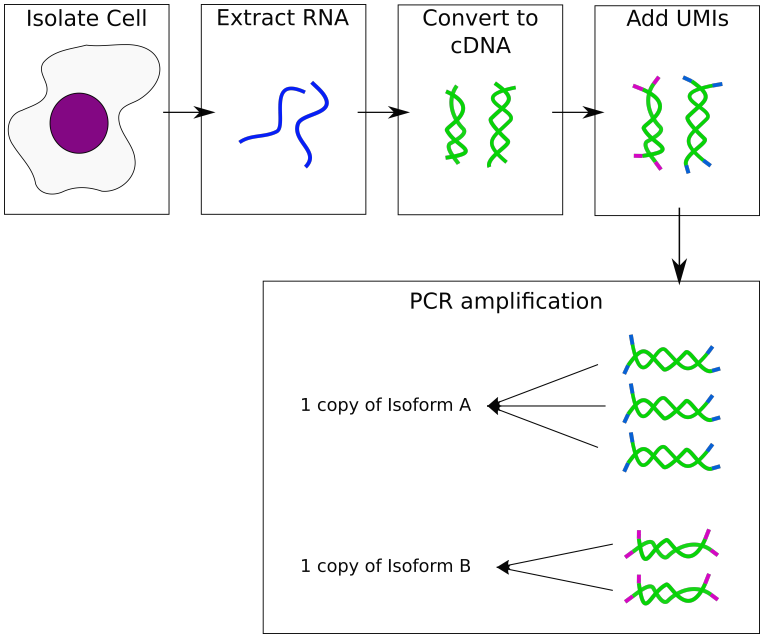


Figure 1.7: Schematic illustrating how addition of UMIs enables us to correct for PCR amplification bias.

are 8 bases or longer (Islam et al., 2014), these errors are usually relatively minor compared to the bias that would be introduced by not correcting for PCR amplification. Methods exist that attempt to correct for UMI errors (Petukhov et al., 2018; Smith et al., 2017). Unfortunately, UMIs are not compatible with all scRNA-seq protocols, and in particular are not compatible with full-length protocols such as SMART-seq2. Given full-length protocols are likely to be best suited to studying alternative splicing, that we are unable to effectively correct for PCR amplification bias in these protocols is concerning.

## 1.11 Previous attempts to study alternative splicing using scRNA-seq

In the first published scRNA-seq study, Tang et al. noted that 8-19% of known genes with two or more isoforms expressed at least two isoforms in individual cells (Tang

et al., 2009). There has therefore been an interest in studying alternative splicing in individual cells since the first days of scRNA-seq. A common approach taken in later studies was to identify genes expressing multiple isoforms in bulk RNA-seq data and to ask how many isoforms were detected from these genes in individual cells using matched scRNA-seq data. These studies found that typically, only one or a few isoforms were detected in individual cells using scRNA-seq, even if multiple isoforms were detected using bulk RNA-seq (Shalek et al., 2013; Zhao et al., 2016; Marinov et al., 2014; Song et al., 2017). However, these studies typically did not consider dropouts, or if they did attempt to correct for dropouts used approximations that now seem inappropriate, given the field's greater understanding of dropout modelling (Marinov et al., 2014). As dropouts impair our ability to detect expressed isoforms and the goal was to count the number of expressed isoforms in each cell, this raises questions about the accuracy of these studies' findings. In addition, the bioinformatic methods to detect isoforms used in these studies had not been independently benchmarked for their performance on scRNA-seq data, raising another potential confounder in these analyses.

In addition to scRNA-seq studies attempting to answer fundamental molecular biology questions about alternative splicing, there have been a small number of studies where scRNA-seq has been used to study alternative splicing to answer questions from other areas of biology. For example, a recent study of a neuronal scRNA-seq dataset investigated the splicing patterns of neurexins, synaptic organisers which have thousands of isoforms, and found that developmentally related cell types had shared patterns of neurexin isoform expression (Lukacsovich et al., 2019). However, the overall number of publications investigating alternative splicing using scRNA-seq is low. This is likely to reflect uncertainty over what confounders are present when trying to study alternative splicing using scRNA-seq, and how best to correct for these confounders.

## 1.12 Approaches for studying alternative splicing with RNA-seq data

It is rare to study alternative splicing using scRNA-seq data, but far more common to study splicing using bulk RNA-seq data. Indeed, a large number of software tools have been developed to enable splicing analyses using bulk RNA-seq data (Li and Dewey, 2011; Roberts and Pachter, 2013; Bray et al., 2016; Patro et al., 2017, 2014; Trapnell et al., 2010). There are two common approaches to studying alternative splicing using bulk RNA-seq data. The first is isoform quantification, in which the goal is to determine the magnitude of expression of splice isoforms. RSEM, Salmon and Kallisto are examples of popular isoform quantification software tools (Li and Dewey, 2011; Bray et al., 2016; Patro et al., 2017). The second approach is sometimes described as an 'exon centric' approach. Instead of attempting to quantify the expression of entire isoforms, a ratio or percentage is calculated for each exon, based on the percentage of reads spanning the exon in which the exon is spliced in versus the percentage of reads spanning the exon in which the exon is spliced out. MISO is a popular example of the 'exon centric' approach (Katz et al., 2010). In my thesis, I have exclusively considered isoform quantification based approaches. My main reason for doing this is that whilst it can be interesting to find the ratio at which an exon is spliced in or out, it would usually be more interesting to know unambiguously which isoforms were produced from the parent gene. From a basic biology and disease perspective, we are often most interested in alternative splicing as a mechanism for generating multiple protein structures from a single gene. If we know which isoforms were produced, we have some chance of inferring which proteins might have been translated. If we only have information at the level of exons, it can be far harder to infer protein information. Therefore, I focus on isoform quantification in my feasibility assessment.

It is recognised that isoform quantification is a hard problem (Garber et al., 2011; Finotello and Di Camillo, 2015; Zhang et al., 2017). Despite a range of strategies having been developed to quantify isoforms (Li and Dewey, 2011; Roberts and Pachter, 2013; Bray et al., 2016; Patro et al., 2017, 2014; Trapnell et al., 2010; Huang and

Sanguinetti, 2017), in benchmarks no strategy is found to perform perfectly (Germain et al., 2016; Teng et al., 2016). Previous scRNA-seq studies have detected one or a small number of isoforms in individual cells for most genes, even if multiple isoforms were detected in matched bulk RNA-seq data (Shalek et al., 2013; Zhao et al., 2016; Marinov et al., 2014; Song et al., 2017). If these observations are accurate, they suggest that isoform quantification may be simpler using scRNA-seq compared to bulk RNA-seq data due to fewer multi-mapping reads. However, at present most scRNA-seq publications quantify reads at a gene rather than an isoform level, probably due to uncertainty over best practices when quantifying at an isoform level. Although many isoform quantification tools are available for bulk RNA-seq, when I began my thesis it was unclear whether these tools would perform well when run on scRNA-seq. For tasks such as normalisation and network modelling, it has been found that methods designed for bulk RNA-seq do not give accurate results when run on scRNA-seq (Vallejos et al., 2017; Chen and Mar, 2018). Therefore, there was a possibility that new software would need to be developed to enable accurate isoform quantification using scRNA-seq.

Although there is currently no perfect strategy for quantifying isoforms (Germain et al., 2016; Teng et al., 2016), existing methods are considered to be sufficiently accurate to enable many studies to analyse alternative splicing using bulk RNA-seq. An important observation from bulk RNA-seq studies is that most genes produce a 'major', more highly expressed isoform and one or more 'minor', less highly expressed isoforms (Wang et al., 2008; Gonzàlez-Porta et al., 2013). How splicing is regulated in individual cells to generate this pattern at the tissue level is not well understood. One hypothesis is that cells exclusively express either the major or the minor isoform, and that more cells express the major isoform than the minor isoform. A second hypothesis is that all cells express both isoforms, and transcribe more copies of the major than the minor isoform. Theoretically, scRNA-seq could enable us to determine the extent to which each of these hypotheses are true by resolving which isoforms are present in individual cells. However, such an approach would rely upon it being possible to accurately resolve isoforms in individual cells using scRNA-seq.

## 1.13 smFISH as an orthogonal approach for studying alternative splicing at a cellular resolution

The main orthogonal approach to scRNA-seq for studying splicing in individual cells is smFISH. In smFISH approaches, cells are fixed and permeabilised. The cells are then hybridised with multiple short fluorescently labeled DNA probes (Femino et al., 1998; Haimovich and Gerst, 2018). The probes have complementary sequences to the targeted RNA species, and together the probes typically span the length of the target transcript. In the permeabilised cell, the probes preferentially hybridise to the target RNA molecule, generating spots of high intensity fluorescence corresponding to target RNA molecules. An advantage of smFISH based approaches is that spatial information (ie. the location of transcripts within the cell) can be obtained, whereas this is not possible using scRNA-seq.

There are two main challenges for using smFISH to study alternative splicing. The first is that it is technically challenging to design fluorescent probes that would enable smFISH to resolve between highly similar isoforms. Previous smFISH studies which have resolved isoforms from the same parent gene have typically used isoforms with large unique regions to overcome this problem (Ciolli Mattioli et al., 2019; Waks et al., 2011; Velten et al., 2015). A variant on smFISH has been developed which enables smFISH to resolve between transcripts which only differ at a single nucleotide variant (SNV) (Levesque et al., 2013). Whether this or a related approach could be adapted to resolve between highly similar isoforms remains to be seen, to the best of my knowledge it has never been attempted.

The second challenge for studying alternative splicing using smFISH is throughput. Traditionally smFISH has been a low throughput technology. Consequently, in previous smFISH experiments investigating how many isoforms are produced per gene per cell, only a handful of genes were investigated (Ciolli Mattioli et al., 2019; Waks et al., 2011; Velten et al., 2015). Whilst these experiments delivered splicing insights for the particular genes investigated, they did not deliver much general in-

sight into the cellular process of alternative splicing across the entire transcriptome. In this respect, using scRNA-seq to study splicing delivers a major advantage over smFISH based approaches. However, the throughput of smFISH based approaches has improved in recent years (Eng et al., 2019; Moffitt et al., 2016). If the throughput continues to improve, it may one day become possible to validate transcriptome wide predictions from RNA-seq experiments using smFISH data.

## 1.14 Is it possible to study alternative splicing using scRNA-seq?

If we are not concerned about the accuracy of our measurements, it obviously is possible to 'study' alternative splicing using scRNA-seq in the sense that there is nothing to stop us running scRNA-seq data through an alternative splicing bioinformatics pipeline. However, as researchers we are hopefully concerned about whether the experiments we perform produce meaningful results. I consider a meaningful result to be a result for which we either have high confidence that the result is correct, and/or for which we have a very good estimate of the error. If the error rate is very high, it may still be challenging to draw biologically valid conclusions from our data, even if we understand the errors in our data very well.

At the start of my PhD, I consider that the following questions were unanswered:

1. Do isoform quantification software tools designed for bulk RNA-seq give accurate measurements when run on scRNA-seq?

2. What impact does the number of cells sequenced and the number of reads sequenced per cell have on isoform quantification in scRNA-seq?

3. To what extent do isoform quantification errors confound alternative splicing analyses in single-cell RNA-seq?

4. To what extent do dropouts confound alternative splicing analyses in scRNA-seq?

5. Could isoform choice within cells confound alternative splicing analyses in scRNA-seq?

6. To what extent is it possible to distinguish between biological signal and technical noise when studying alternative splicing using scRNA-seq?

As a consequence of not knowing the answer to these questions, we had no clear idea what errors we might encounter when studying alternative splicing using scRNA-seq. Of course, this does not mean it is impossible to perform an alternative splicing analysis with scRNA-seq data, as many have previously done. But it does mean that after running such an analysis, it is hard to say how much confidence we can have in the results of the analysis, because we do not understand what errors might be present.

To address our lack of knowledge about the errors present in scRNA-seq, I first performed a benchmark of isoform quantification softwares, which addressed questions 1 and 2. This work is presented in chapter 2. I next attempted to address a series of biological alternative splicing questions using scRNA-seq, but found that uncertainty over the degree of technical noise in scRNA-seq made interpreting my results in a biological context impossible. These results are presented in chapter 3. To address the challenges I faced in chapter 3, I used a novel simulation approach to investigate the degree to which technical noise might be confounding my splicing analyses in chapter 4. In chapter 4, I address questions 3-6, and thus answer the over-arching question which motivated my thesis: Is it feasible to study alternative splicing using scRNA-seq?

# 2

# Simulation Based Benchmarking of Isoform Quantification Using scRNA-seq.

*The truth is rarely pure and never simple.*
– Oscar Wilde, *The Importance of Being Earnest*(Wilde, 1895)

## Preface

The overall goal of the work presented in this chapter is to answer the following questions:

1. Do isoform quantification software tools designed for bulk RNA-seq give accurate measurements when run on scRNA-seq?

2. What impact does the number of cells sequenced and the number of reads sequenced per cell have on isoform quantification in scRNA-seq?

I address these questions by carrying out a simulation based benchmark of isoform quantification using scRNA-seq. I find that four of the five isoform quantification

tools evaluated in my benchmark perform almost as well when run on scRNA-seq as when run on bulk RNA-seq, thus answering question 1. I find that the number of cells sequenced has no impact on the performance of a popular isoform quantification tool called Salmon. Over a range of read depths per cells, the performance of Salmon peaks at 1-2 million reads per cell, implying that there may be an optimum read depth for performing isoform quantification using scRNA-seq.

The work presented in this chapter has been published, consequently some passages have been quoted verbatim from the following sources: (Westoby et al., 2018a,b). Additionally, some figures have been reproduced from the aforementioned sources.

## 2.1 Introduction

Numerous isoform quantification tools have been developed for bulk RNA-seq (Li and Dewey, 2011; Roberts and Pachter, 2013; Bray et al., 2016; Patro et al., 2017, 2014; Trapnell et al., 2010), however at the start of my thesis, it was not clear whether these tools would perform appropriately when run on scRNA-seq data. scRNA-seq data differs from bulk RNA-seq data in several notable ways. For example, due to the low amount of starting material, there is an increased frequency of dropouts and increased PCR amplification bias in many scRNA-seq protocols relative to bulk RNA-seq (Islam et al., 2014; Kharchenko et al., 2014). It was not obvious what effect these technical factors, and others, might have on the performance of isoform quantification tools.

In addition to general technical concerns, an interesting question was whether the performance of isoform quantification tools might differ depending on the library preparation protocol used to generate the scRNA-seq data. A wide range of library preparation protocols have been developed for scRNA-seq (Hashimshony et al., 2012, 2016; Macosko et al., 2015; Klein et al., 2015; Jaitin et al., 2014; Gierahn et al., 2017; Picelli et al., 2014; Ramsköld et al., 2012; Islam et al., 2011), some of which are likely to be more appropriate for isoform quantification than others. For example, one way in which library preparation protocols could differ in their suitability for isoform quantification is in their degree of gene length bias, which has been shown

to be greater for full length transcript protocols compared with UMI based protocols (Phipson et al., 2017). An understanding of which library preparation protocols generate data suitable for isoform quantification and which library preparation protocols do not would allow researchers to better design experiments to suit their needs.

There is currently a trade-off in scRNA-seq between the number of cells sequenced and the number of reads sequenced per cell (Bacher and Kendziorski, 2016). A pertinent question is whether the number of cells sequenced or the number of reads sequenced per cell impacts on the performance of isoform quantification tools when run on scRNA-seq. For bulk RNA-seq, a wide range of read numbers have been sequenced depending on the desired accuracy of quantification and whether it is desirable to detect and quantify lowly expressed transcripts (Conesa et al., 2016). However, it has been recognised that whilst sequencing at higher read depths can increase the accuracy of quantification, in the context of differential expression, a higher number of reads can also increase the number of false positives if not corrected for (Tarazona et al., 2011). For scRNA-seq, a decision on how many reads to sequence per cell is often driven by multiple factors, including how many cells should be sequenced and whether the goal of the experiment is to find detailed information on gene expression in each cell or to identify sub-populations of cells by clustering or pseudotime analysis (Haque et al., 2017). A recent study suggests that if the goal is clustering cells, it is more beneficial to sequence more cells than to sequence fewer cells slightly more deeply (Svensson et al., 2019). Multiple studies have found that the number of genes detected reaches saturation for current library preparation protocols at around 1 million reads per cell (Wu et al., 2014; Ziegenhain et al., 2017), suggesting that for gene detection there is little purpose in further increasing sequencing depth. Whether increased sequence depth would improve the accuracy of isoform quantification is not known.

To assess isoform quantification for scRNA-seq, I present a simulation based benchmarking study using data generated from three different scRNA-seq projects. Whilst benchmarking studies have been performed previously for bulk RNA-seq (Germain et al., 2016; Teng et al., 2016), to the best of my knowledge this is the first benchmark of isoform quantification performed for scRNA-seq. I evaluated the over-

all accuracy of different isoform quantification methods when applied to scRNA-seq, and I also specifically studied the impact of library preparation protocol and dropouts. I tested five popular isoform quantification tools on simulated scRNA-seq data based on three publicly available scRNA-seq datasets produced using different library preparation protocols and cell types. Unless otherwise stated, in all of my simulations in this chapter, all of the isoforms in the Ensembl 89 mouse transcriptome were considered (approximately 120,000 transcripts). The entire transcriptome was considered because researchers are commonly interested in a variety of different isoforms, from highly expressed isoforms that are detected in most cells to lowly expressed isoforms that are only rarely detected. With the exception of eXpress, performance was generally good for simulated data based on SMARTer and SMART-seq2 (Picelli et al., 2014) data. Compared to bulk RNA-seq, isoform quantification was only slightly worse for SMARTer and SMART-seq2 data, suggesting that it is appropriate to use these methods for full-transcript single-cell data.

## 2.2 Results

### 2.2.1 The performance of isoform quantification tools was generally good and consistent across two different simulation methods.

The first dataset considered in this benchmark consisted of 96 mouse quiescent B lymphocytes collected as part of the BLUEPRINT epigenome project (Adams et al., 2012) (GEO accession code GSE94676). The SMARTer library preparation protocol was used to collect this dataset, which has been shown to have a degree of 3' coverage bias (Wu et al., 2014). On average, just over 2.7 million reads had been sequenced per B lymphocyte.

To perform the benchmark, simulated data was generated from the selected cells using two simulation methods. The first simulation method used was RSEM (Li and Dewey, 2011) (see Methods chapter for details). RSEM is an isoform quantification

tool which uses a generative model and expectation maximization to estimate isoform expression. In addition, RSEM is capable of simulating reads using its generative model and input values for the latent variables in the model, which can be estimated during isoform quantification. An important reason for selecting RSEM to perform the simulations is that during the simulation process, RSEM records where each simulated read originated in the transcriptome. Consequently, it is known how highly expressed each isoform is in the simulated data. This will be referred to as the 'ground truth.' Knowing the ground truth allows us to benchmark expression estimates from isoform quantification tools using the simulated data.

The second simulation method relied on two tools, Splatter (Zappia et al., 2017b) and Polyester (Frazee et al., 2015). The methodology used to generate simulated data is illustrated in Figure 2.1. Splatter is a simulation tool which takes an expression matrix of counts from an scRNA-seq experiment as input and gives a simulated expression matrix of counts as output. Splatter was used to simulate counts data based on an expression matrix of counts from the BLUEPRINT B lymphocytes generated by isoform quantification tool Kallisto (Bray et al., 2016). The output of Splatter is a gene count expression matrix, where the columns are cells and the rows are non-specific gene names (e.g., 'Gene1', 'Gene2', 'Gene3'). Polyester was then used to simulate one read per count in the Splatter gene count expression matrix. Since the exact origin in the transcriptome is not known from Splatter, Polyester generated simulated reads using a transcriptome consisting of the isoforms called as expressed by Kallisto in at least one cell. The rownames of the Splatter count matrix were updated to reflect the isoforms simulated by Polyester. The Splatter count matrix was then converted to a matrix of TPM values, which were used as the 'ground truth'.

The RSEM- and Splatter- and Polyester-simulated reads data was then given as input to RSEM, eXpress (Roberts and Pachter, 2013), Kallisto, Salmon (Patro et al., 2017), and Sailfish (Patro et al., 2014). The isoform quantification tools provide two useful pieces of information for each isoform—whether it is expressed and its expression level. To quantify the ability of each method to detect the presence of an isoform, the precision and recall were calculated. In this context, the precision is

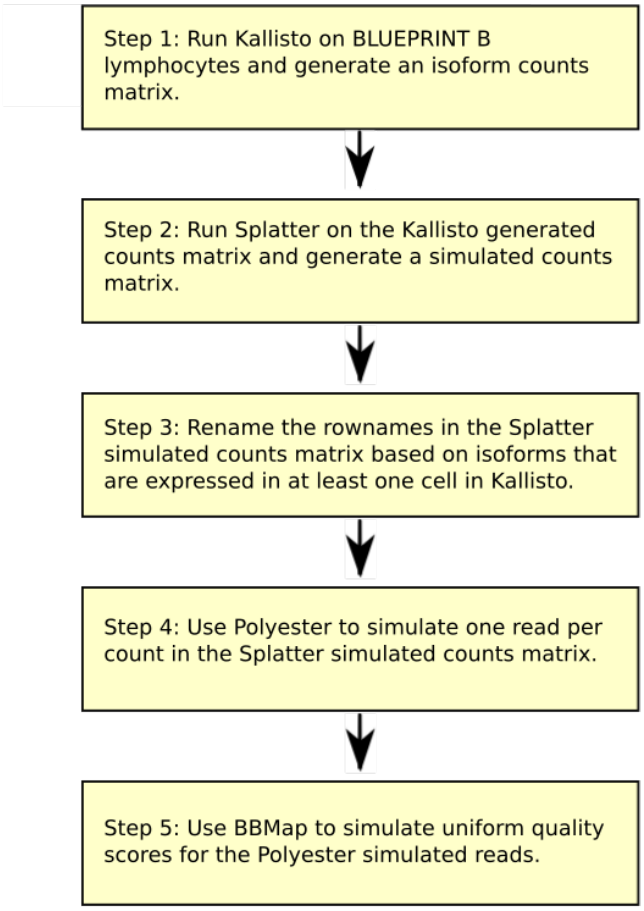


Figure 2.1: Flowchart showing methodology for generating Splatter- and Polyester-simulated data

the fraction of isoforms predicted to be expressed by each tool which are expressed in the ground truth. The recall is the fraction of isoforms expressed in the ground truth which are predicted to be expressed using the tool. For a single overall quality score, I used the F1 score, which is defined as the harmonic mean of precision and recall.

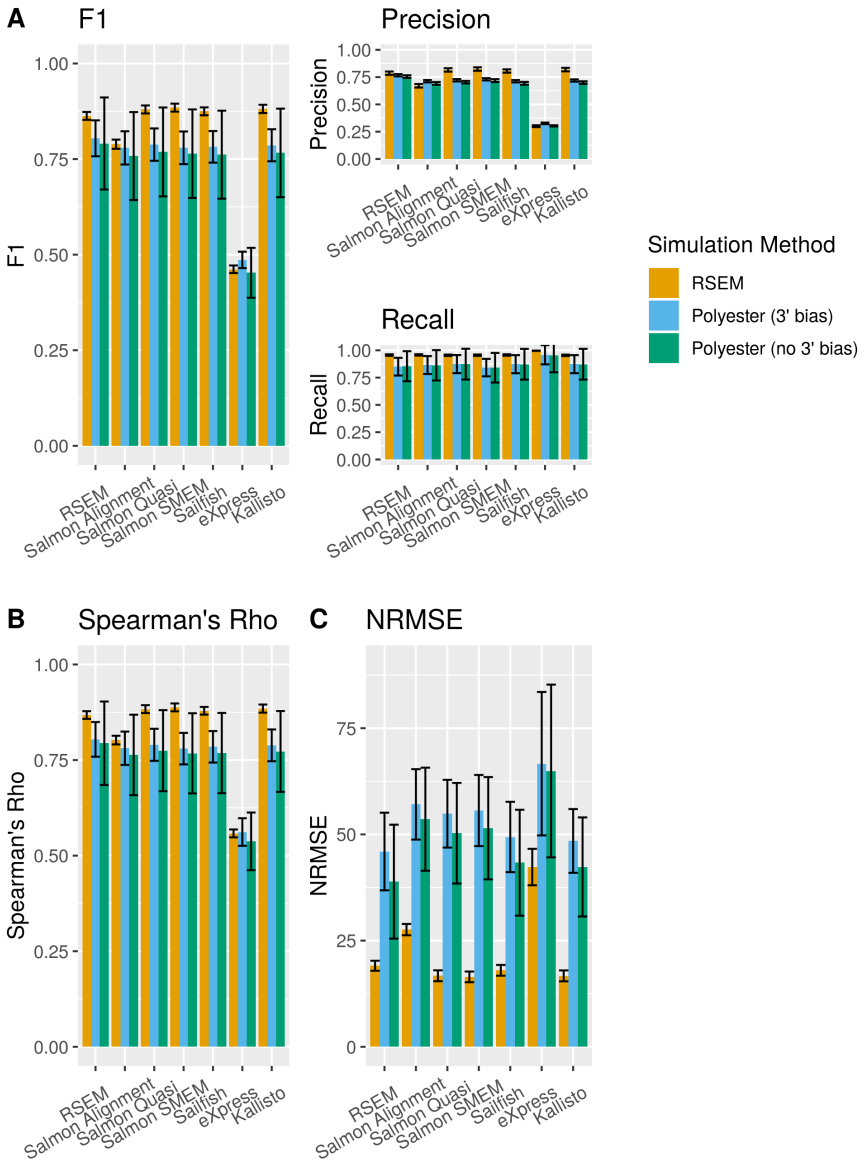


Figure 2.2: Performance statistics for each isoform quantification tool for the BLUEPRINT simulations. The yellow bars represent RSEM simulations, the blue bars represent Splatter and Polyester simulations with 3' coverage bias and the green bars represent Splatter and Polyester simulations with no coverage bias. The bars represent the average performance across all simulated cells, the error bar limits are defined by the standard deviation. **A** F1 score, precision and recall of isoform detection. The F1 score is the harmonic mean of the precision and recall. The precision is the proportion of the isoforms predicted to be expressed by an isoform quantification tool which are expressed. The recall is the proportion of expressed isoforms which are predicted to be expressed by the isoform quantification tool. **B** Spearman's rho. **C** Normalised root mean square error (NRMSE)

Salmon can be run in three modes—an alignment-based mode, in which aligned reads are taken as input, or one of two alignment free modes (a quasi mode or an SMEM mode). The performance of all three modes was evaluated in this benchmark. For most isoform quantification tools, the mean F1 score was remarkably similar and in the range of 0.777–0.888. The exception was eXpress, which had a slightly higher recall but a much lower precision than other tools, and consequently had the lowest mean F1 score (between 0.463 and 0.492 depending on the simulation method) (Figure 2.2A). The mean F1 scores, precisions, and recalls calculated for each of these tools were similar regardless of whether RSEM or Splatter and Polyester were used to generate the simulated data. The statistics were not dramatically altered when Polyester simulated reads using a 3' coverage bias model compared to when Polyester simulated reads uniformly across transcript length. However, as the Polyester 3' coverage bias model is not based on single-cell RNA-seq data, care needs to be taken when interpreting this result.

In addition to determining whether an isoform is expressed, it is often of interest to estimate isoform abundance. To evaluate how well isoform quantification tools perform this task, two measures were considered—Spearman's rho and the normalised root mean square error (NRMSE) (Figure 2.2B, C). Spearman's rho gives a measure of how monotonic the relationship between the ground truth expression and each tool's expression estimates is, while the NRMSE gives a measure of the extent to which the relationship deviates from a one to one linear relationship (see Methods chapter for details on how the NRMSE was calculated).

Consistent with the results for isoform detection, mean Spearman's rho was similar between isoform quantification tools and simulation methods and in the range 0.782–0.891. The exception was eXpress, which had much lower mean Spearman's rho than the other tools with values from 0.550 to 0.574. eXpress also performed poorly relative to the other isoform quantification tools when considering the NRMSE. Although the overall pattern of NRMSE results was similar for both simulation methods, the NRMSE was consistently far higher for the Splatter and Polyester simulations compared to the RSEM simulations. One possible explanation is that the difference in the NRMSE is due to a small number of outliers. However,

this did not appear to be the case (see Figure 2.3). Another explanation for the difference in the NRMSE could be that the differences are largely driven by differences in the ground truth expression distributions of the RSEM simulations compared to the Splatter and Polyester simulations. Since the NRMSE is proportional to the sum of squared differences between the ground truth and the isoform quantification tool's expression estimates, it is plausible that it will be relatively rare for an unexpressed isoform to have an estimated expression other than zero, but relatively common for an expressed isoform to have an estimated expression that differs from the ground truth expression. I found that the distribution of ground truth expression values differs for each simulation method (see Figure 2.4). Therefore, differences in the ground truth expression distributions seem to be the most likely explanation for the systematic difference in the NRMSE between simulation methods.

The difference in the NRMSE between simulation methods was not the only aspect in which the simulation methods differed. A comparison of the simulated data with the real data was carried out using both a comparison tool included in Splatter and using CountsimQC (Soneson and Robinson, 2018a), a package which facilitates comparison of simulated datasets. The RSEM-simulated data more closely resembled the real data than the Splatter and Polyester-simulated data by a number of metrics, including the sample-sample correlations, the mean-variance relationship, and the relationship between magnitude of expression and fraction of zeros (see Figure 2.5). In contrast, when comparing the simulation tools using gene-level statistics such as the distribution of mean expression, distribution of variance and percentage of zeros per gene, the resemblance between the Splatter/Polyester-simulated data and real data is much closer (see Figure 2.6). I suspect that these differences are because Splatter loses gene names during its simulations. When the Splatter counts matrix was used with Polyester to simulate reads data, I updated the row names to reflect the transcripts simulated by Polyester. Consequently, I would expect there to be little or no relationship between the expression of a given gene in real data and the corresponding Polyester/Splatter-simulated data. Indeed, I find that the correlation between ground truth isoform expression in the Splatter- and Polyester-simulated data and isoform expression estimates generated by running Kallisto on the real

BLUEPRINT B lymphocyte data is very low (see Figure 2.7). In contrast, the correlation between ground truth expression in the RSEM simulations and Kallisto expression estimates in the real data was much higher.

The difference in the correlation between the real data and the RSEM simulations compared with the Splatter/Polyester simulations is probably due to a core difference between the simulation methodologies. RSEM keeps isoform names throughout its simulations, and bases the number of reads it simulates for each isoform in part on how many reads it detected for that isoform in the real data. Therefore, we would expect an isoform's RSEM simulated expression level to correlate with its expression level in the real data. In contrast, in the Splatter/Polyester simulations, isoform names are not kept. Splatter simulates isoform expression levels based on the global distribution of isoform expression, in addition to other factors. Splatter does not name its isoforms based on the original isoform names, but renames them 'Gene1', 'Gene2', etc. The Splatter counts are then used by Polyester to simulate reads, but the mapping between a given isoform's expression in the real data and in the simulated data is lost due to the loss of isoform names.

Therefore, different transcriptional profile in the Splatter- and Polyester-simulated data compared with the real BLUEPRINT B lymphocyte data is a likely consequence of updating the Splatter gene names to reflect the transcriptome used in the Polyester simulations (Step 3 in Figure 2.1). An additional potential issue with this step in my methodology is that factors which would normally impact on expression estimates, such as gene length, GC content, and secondary structure, are not considered during my simulation protocol. The relationships between these features and expression estimates in my simulated data are unlikely to match the real data. Based on these limitations and my findings above, I concluded that the RSEM simulations resembled the real data more closely than the Splatter and Polyester simulations. I suspect that this occurs due to the loss of gene labels during the Splatter simulations and subsequent reassignment during the Polyester simulations, leading to a radically different transcriptional profile in the Splatter/Polyester-simulated data. Consequently, for the rest of this chapter, all data was simulated using RSEM. Despite the differences between the RSEM and Splatter and Polyester simulations, the results of the bench-

mark were remarkably consistent. This suggests that the findings in this benchmark are robust to some differences between datasets, including dramatic changes in the transcriptional profile.

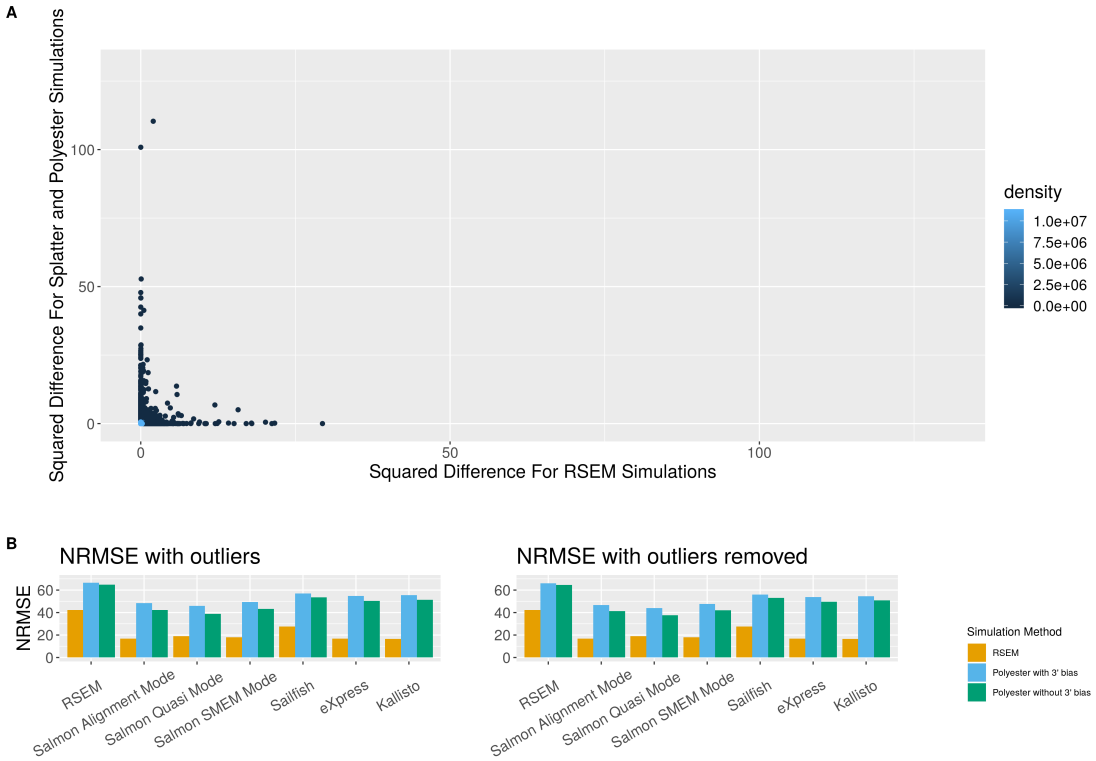


Figure 2.3: The difference in the NRMSE between RSEM and Splatter and Polyester simulations could not be explained by outliers. **A** For each isoform, the mean squared difference between RSEM's expression estimates and the ground truth was calculated for the Splatter and Polyester simulations and for the RSEM simulations. Points are coloured by density.**B** The NRMSE when outliers were and were not removed. Based on **A**, outliers were defined as isoforms with a mean squared difference greater than 30 in the Splatter and Polyester simulations.

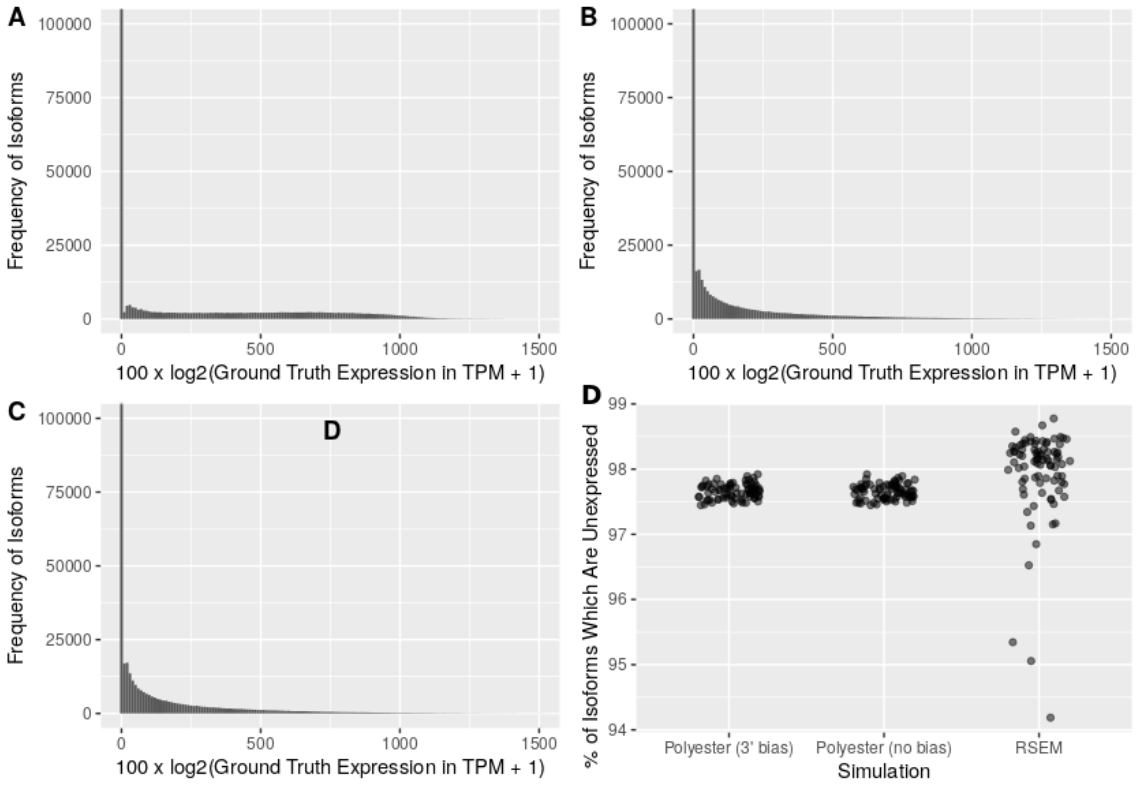


Figure 2.4: Histograms of ground truth expressions values for different simulation methods. **A** Histogram of ground truth expression values for RSEM simulations. **B** Histogram of ground truth expression values for Splatter and Polyester simulations with 3' coverage bias. **C** Histogram of ground truth expression values for Splatter and Polyester simulations with no coverage bias. **D** Percentage of isoforms which are unexpressed (ie. have zero expression) in RSEM and Splatter and Polyester simulations. Each point represents one simulated cell.

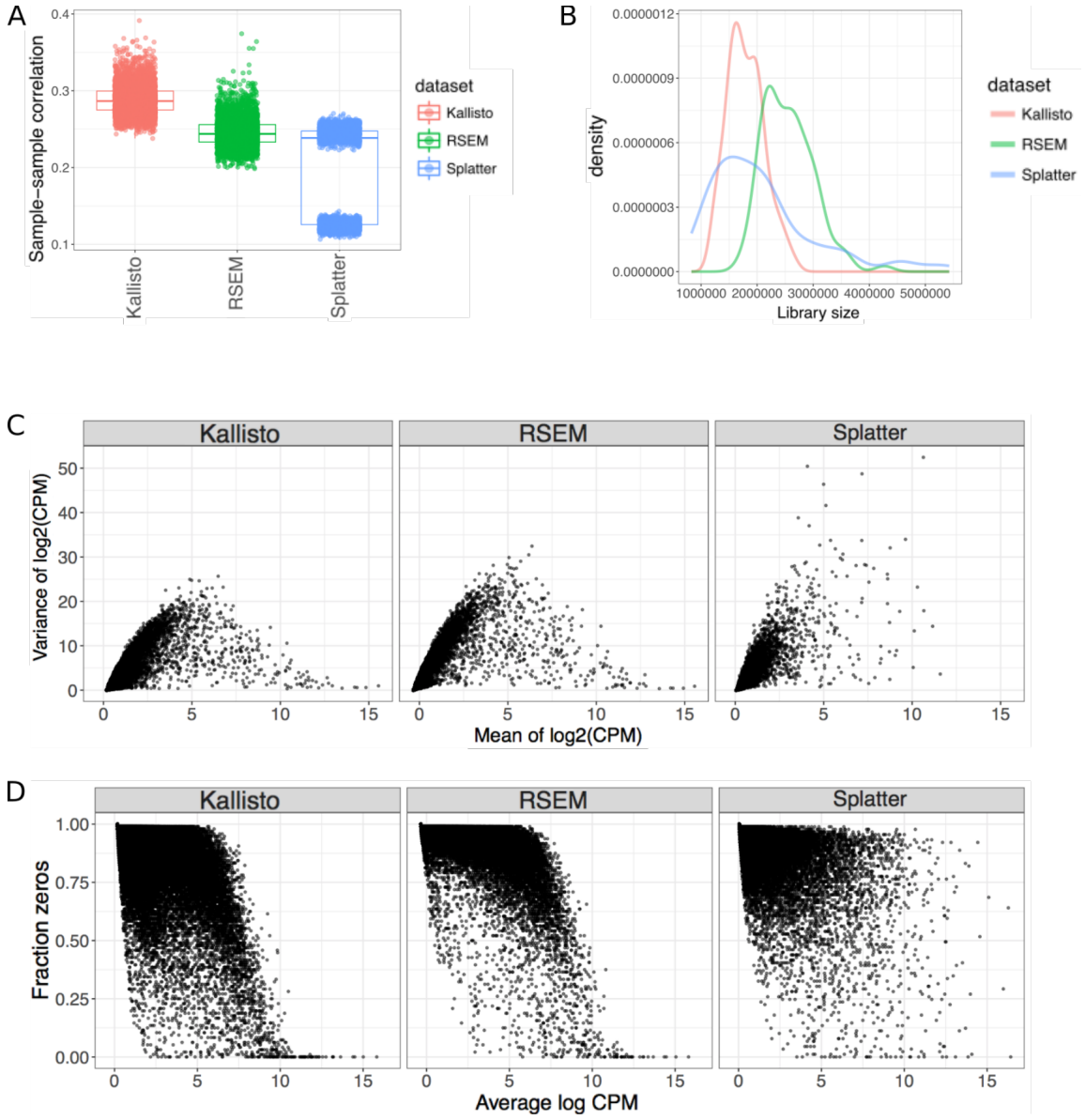


Figure 2.5: A comparison of the RSEM and Splatter simulations with the real BLUEPRINT data. CountsimQC(Soneson and Robinson, 2018a) was used to generate these figures, using expression estimates generated by running Kallisto(Bray et al., 2016) on the real BLUEPRINT B lymphocytes (red), ground truth expression values from the RSEM simulated data (green) and ground truth expression values from the Splatter and Polyester simulated date (blue). **A** Boxplots of sample-sample correlations. Each point represents the Spearman correlation coefficient between two cells. **B** Frequency density plot of library sizes. **C** Scatter plots of the mean-variance relationship for log2(CPM). **D** Scatter plots showing the relationship between the fraction of zeros and the average log CPM.

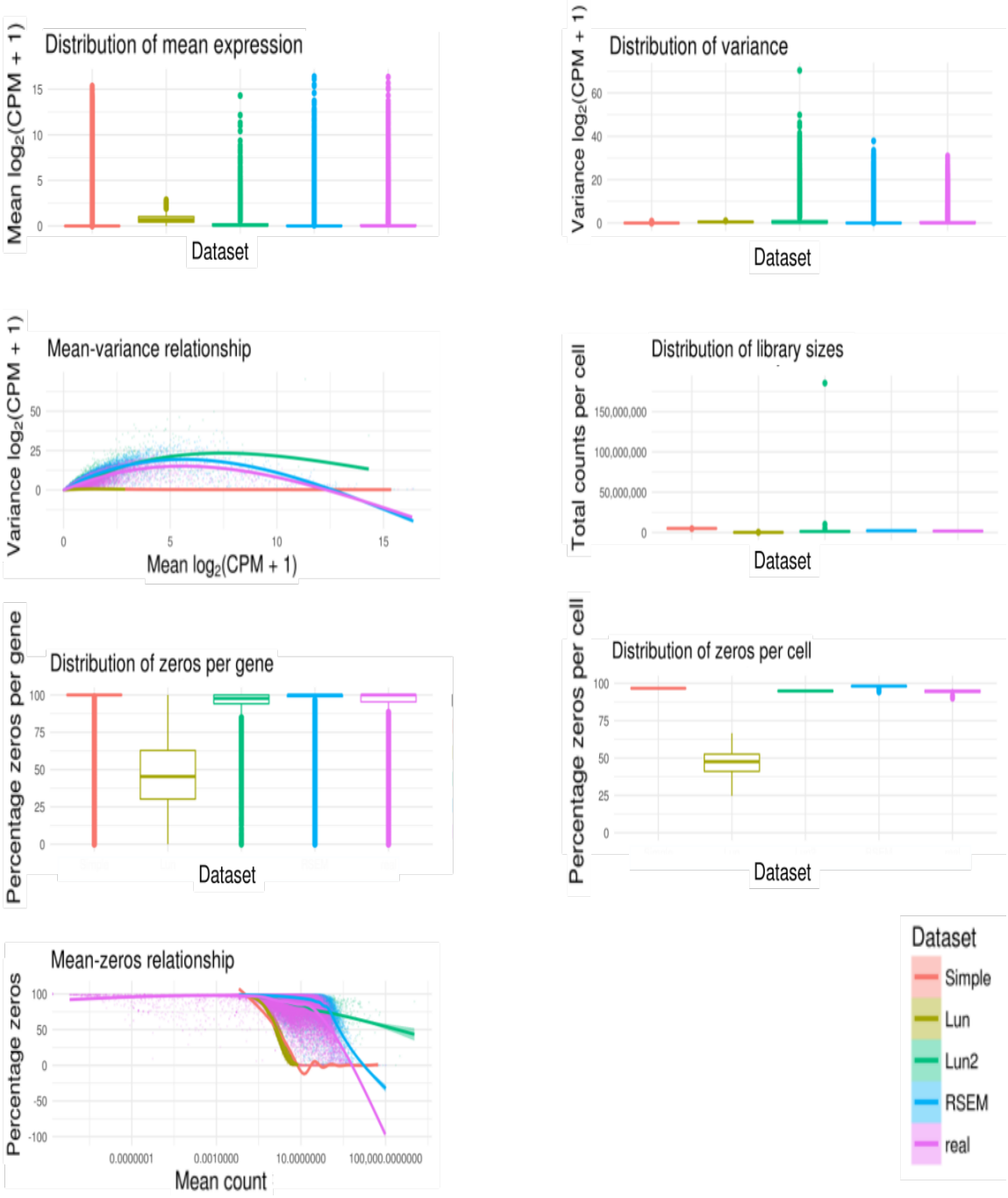


Figure 2.6: Plots showing characteristics of the different simulation methods included within Splatter, compared with the RSEM simulations and the real data. Based on these plots, the Lun2 method was selected for use in the rest of this chapter.

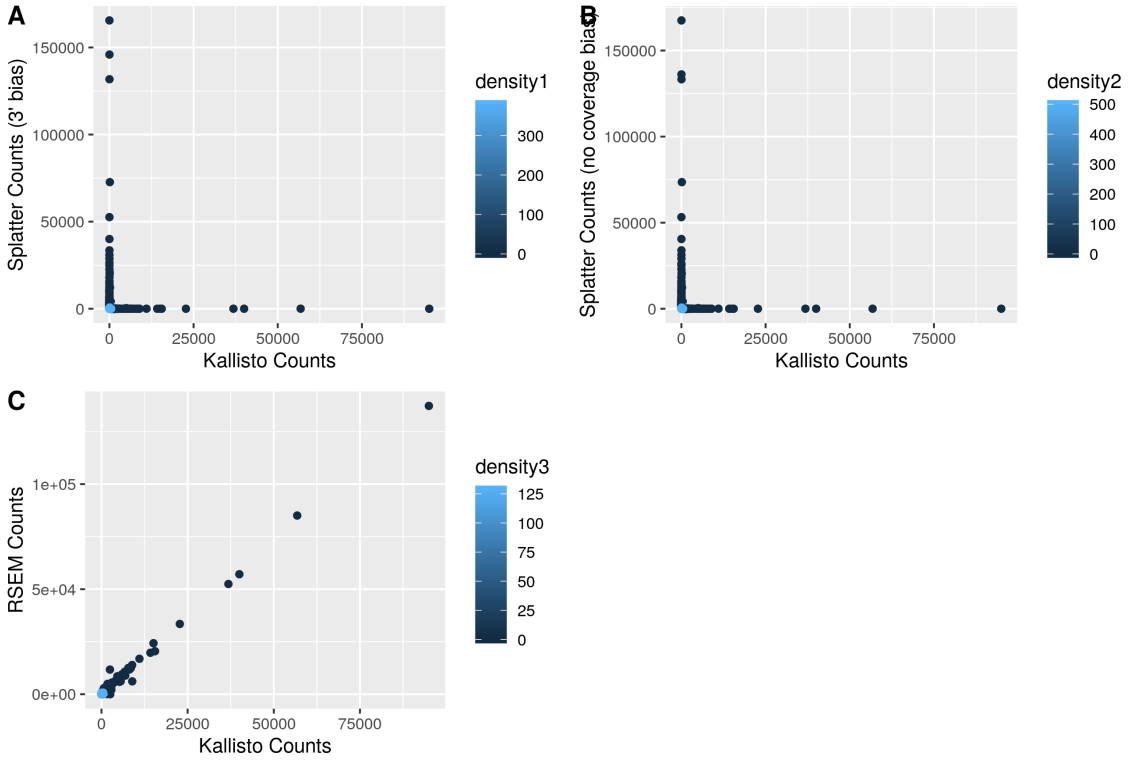


Figure 2.7: The relationship between expression estimates generated by running Kallisto on the real BLUEPRINT B lymphocytes and the ground truth expression estimates from simulated data. Points are coloured by density.**A** Relationship between expression estimates generated by running Kallisto on the real BLUEPRINT B lymphocytes and the ground truth expression values from the Splatter and Polyester 3' bias simulated data. **B** Relationship between expression estimates generated by running Kallisto on the real BLUEPRINT B lymphocytes and the ground truth expression values from the Splatter and Polyester simulated data with no coverage bias. **C** Relationship between expression estimates generated by running Kallisto on the real BLUEPRINT B lymphocytes and the ground truth expression values from the RSEM simulated data

### 2.2.2 Isoform quantification tools generally perform well on SMART-seq2 data with high sequencing coverage.

To test whether the results of my benchmark were robust across different datasets, I next considered a mouse embryonic stem cell (mESC) dataset published by Kolodziejczyk et al. (Kolodziejczyk et al., 2015). On average, over 7 million reads were sequenced per cell in this dataset, considerably more than in the BLUEPRINT dataset. Intuitively, it seems likely that sequencing more reads per cell should lead to improved isoform quantification. However, sequencing more reads per cell is expensive and may come at the cost of being unable to sequence as many cells. Therefore, determining whether sequencing a higher number of reads per cell improves isoform quantification is likely to be of interest to many researchers.

From the Kolodziejczyk et al. dataset, 271 mESCs grown in standard 2i media + LIF which passed quality control were used for the benchmark (see Methods chapter). This dataset should therefore give a good indication of the performance of isoform quantification tools when there are a high number of reads per cell but a relatively low number of cells. In addition, this dataset has uniform coverage of transcripts, as it was sequenced using the SMART-seq2 protocol (Picelli et al., 2014).

To perform the benchmark, simulated data was generated as described previously from the selected cells from Kolodziejczyk et al. (see Methods chapter for details). The simulated reads data were then given as input to RSEM, eXpress, Kallisto, Salmon, and Sailfish. The highest F1 score was achieved by Salmon run in SMEM mode (0.889), with RSEM, Salmon run in quasi mode, Sailfish, and Kallisto also achieving mean F1 scores greater than 0.85 (Figure 2.8A). Again, eXpress performed most poorly by a substantial margin, with a mean F1 score of 0.548, and again, eXpress had a higher mean recall (0.997) but a much lower mean precision (0.378) than other tools. It seems likely that eXpress's low precision is due to it being too liberal when calling isoforms as expressed. The average number of isoforms called as expressed per cell was twice as high for eXpress, which called an average of 41,372 isoforms as expressed per cell, as for any other tool. The other isoform quantification tools had high mean recalls between 0.956 and 0.960. In contrast, the highest mean

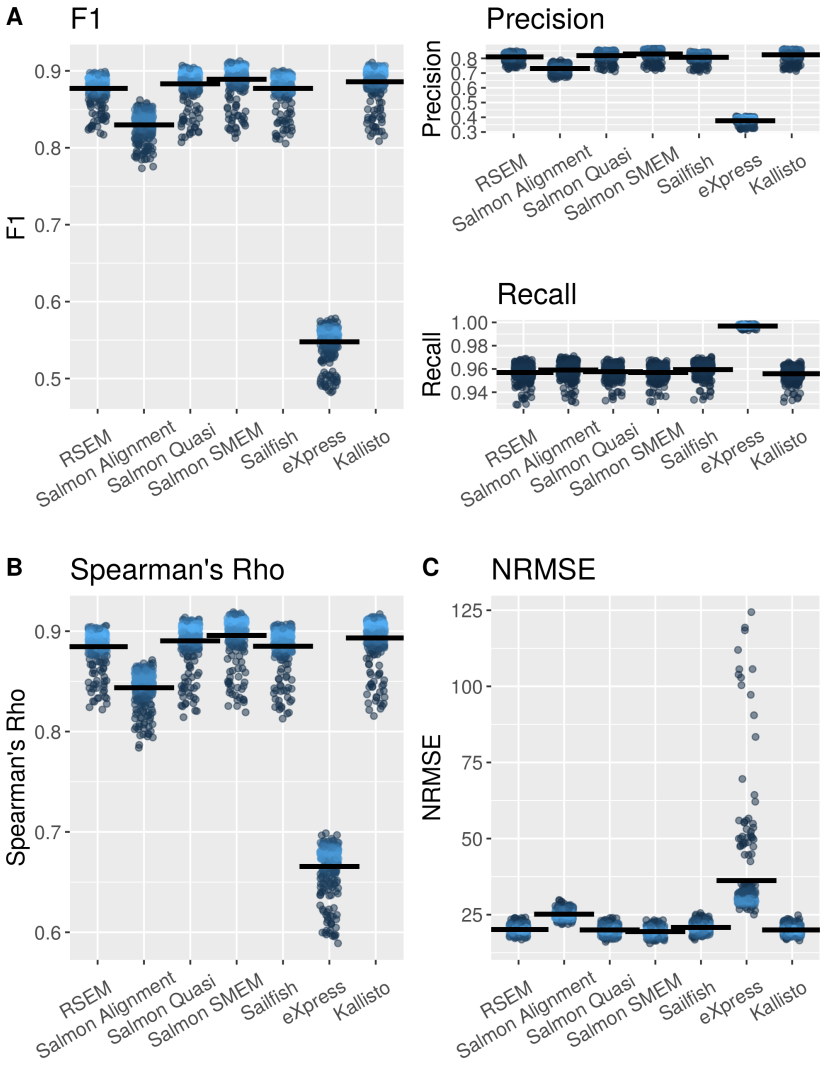


Figure 2.8: Performance statistics for each isoform quantification tool for the Kolodziejczyk et al. ES cell simulations. POints are coloured by density. **A** F1 score and precision and recall of isoform detection. The F1 score is the harmonic mean of the precision and recall. The precision is the proportion of the isoforms predicted to be expressed by an isoform quantification tool which are expressed. The recall is the proportion of expressed isoforms which are predicted to be expressed by the isoform quantification tool. **B** Spearman's rho. **C** Normalised root mean square error (NRMSE)

precision was just 0.831 by Salmon run in SMEM mode, which means that nearly one in six isoforms predicted to be expressed by the best performing tool were not actually expressed. The high recall values achieved by all the tools considered here indicate that the vast majority of isoforms expressed in the simulated data are detected, with the lower precision values being a greater cause for concern. Knowing that an isoform is not expressed can be as important as knowing that an isoform is expressed, especially if that isoform is being used as a marker, for example in clustering analysis. A strategy for improving the detection ability as quantified by the F1 score of isoform quantification tools for scRNA-seq could be to make future tools more conservative when calling isoforms as expressed.

The highest mean value of Spearman's rho was obtained by Salmon run in SMEM mode (0.896), with Salmon run in quasi mode, Kallisto, RSEM, and Sailfish obtaining similar values. The lowest mean value of the NRMSE was also obtained by Salmon run in SMEM mode (19.5), with Salmon run in quasi mode, Kallisto, RSEM, and Sailfish obtaining similar values. Again, of the tools considered, eXpress performed most poorly by a substantial margin.

### 2.2.3 The performance of isoform quantification tools was generally poor using the Drop-seq library preparation method.

Droplet based library preparation methods for scRNA-seq enable tens or hundreds of thousands of cells to be sequenced in a single experiment, but at a relatively low coverage per cell (Macosko et al., 2015; Klein et al., 2015; Gierahn et al., 2017; Zheng et al., 2017). To determine whether a high number of cells can compensate for low sequencing depth, a Drop-seq dataset of retinal bipolar cells published by Shekhar et al. (Shekhar et al., 2016) was considered. Approximately 45,000 cells were sequenced at a median read depth of 8,200 mapped reads per cell in this dataset. From the dataset, 1,000 cells were randomly selected and given as input to RSEM to generate simulated data (see Methods chapter for details). The simulated reads were then given as input to RSEM, eXpress, Kallisto, Salmon and Sailfish as before,

and the performance of these tools for the Drop-seq simulated data was evaluated. With the exception of RSEM, the mean F1 score is far lower for the Shekhar et al. Drop-seq simulated data as compared with the Kolodziejczyk et al. or the BLUEPRINT simulated data (Figure 2.9A). For most tools, this is a consequence of a drop in both the precision and the recall. The mean precision is less than 0.5 for most isoform quantification tools, including those which performed well on the Kolodziejczyk et al. and BLUEPRINT simulated data. Salmon run on SMEM mode, which achieved the highest mean precision on the Kolodziejczyk et al. simulated data, performed particularly poorly, achieving a mean precision of just 0.399. Only RSEM and Kallisto achieved mean precisions greater than 0.5 (0.743 and 0.614 respectively).

This result has important implications for the notion that sequencing a high number of cells could capture the overall transcriptional profile of a population of cells despite a low number of reads per cell. If sequencing a high number of cells could compensate for a low coverage per cell, a very high precision of isoform detection would be required. Without a high precision, attempting to combine data from multiple cells to recapture the population transcriptional profile will result in calling a high number of unexpressed isoforms as expressed, whereas if low read numbers only reduced recall it could be compensated for by combining data across cells. In addition to a low precision and recall, the isoform quantification tools perform relatively poorly on the Shekhar et al. Drop-seq simulated data when Spearman's rho and NRMSE are used as performance metrics (Figure 2.9B, C). Again, Kallisto and RSEM perform relatively well by these metrics compared to the other tools.

The overall picture painted by these results is that a low number of reads per cell reduces the performance of isoform quantification tools, and this cannot be compensated for by sequencing more cells. RSEM appears to perform better than the other isoform quantification tools when run on the Shekhar et al. Drop-seq simulated data, however this result needs to be interpreted with caution. Since RSEM was used to perform the Shekhar et al. Drop-seq simulations, and as it essentially uses the same model to perform isoform quantification and simulations, it is plausible that RSEM's performance is close to optimal when run on its own simulated data. This was also the case for the Kolodziejczyk et al. and some of the BLUEPRINT simulations, but

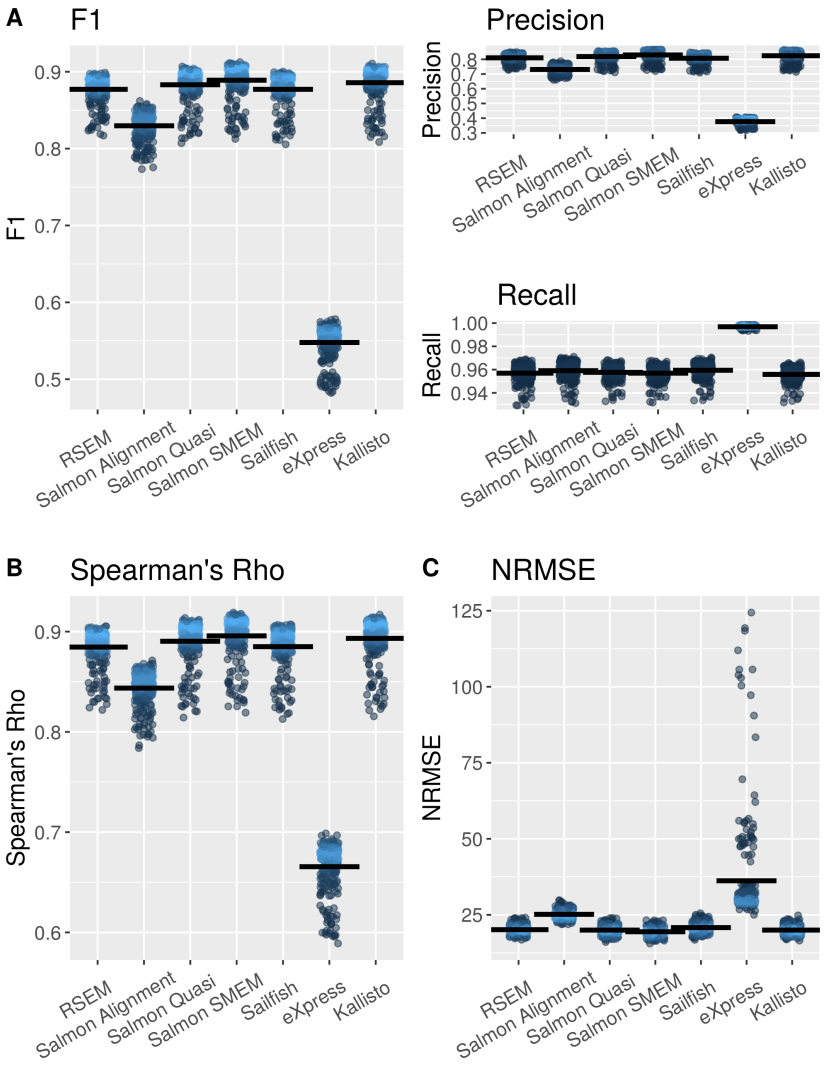


Figure 2.9: Performance statistics for each isoform quantification tool for the Shekhar et al. Drop-seq simulations. Points are coloured by density. **A** F1 score, precision and recall of isoform detection. The F1 score is the harmonic mean of the precision and recall. The precision is the proportion of the isoforms predicted to be expressed by an isoform quantification tool which are expressed. The recall is the proportion of expressed isoforms which are predicted to be expressed by the isoform quantification tool. **B** Spearman's rho. **C** Normalised Root Mean Square Error (NRMSE).

for these simulations the performance of Sailfish, Salmon and Kallisto was not dissimilar to the performance of RSEM. One hypothesis generated from these observations is that on high quality single cell datasets, most isoform quantification tools perform well, meaning that if RSEM is performing optimally, it provides a relatively small advantage. However, on a dataset with a low number of reads per cell, short reads and 3' coverage bias, most tools perform poorly. If RSEM is performing optimally, this may result in a much greater impact on relative performance.

### 2.2.4 The decrease in the performance of isoform quantification using scRNA-seq compared with bulk RNA-seq is generally small

I find that the performance of existing isoform quantification tools is generally good when run on simulated data based on SMART-seq2 and SMARTer scRNA-seq data. I next consider the performance of isoform quantification tools when scRNA-seq data is used compared with bulk RNA-seq data. Although previous benchmarks of isoform quantification have been performed using bulk RNA-seq data (Germain et al., 2016; Teng et al., 2016), a direct comparison with my benchmark is challenging due to differences in the experimental approaches taken. Consequently, it is not possible to say whether any perceived change in the performance of a given tool in my benchmark compared with a bulk RNA-seq benchmark is due to differences in how the benchmark was performed, differences in which statistics were collected, or due to a genuine difference in performance on bulk and single-cell data.

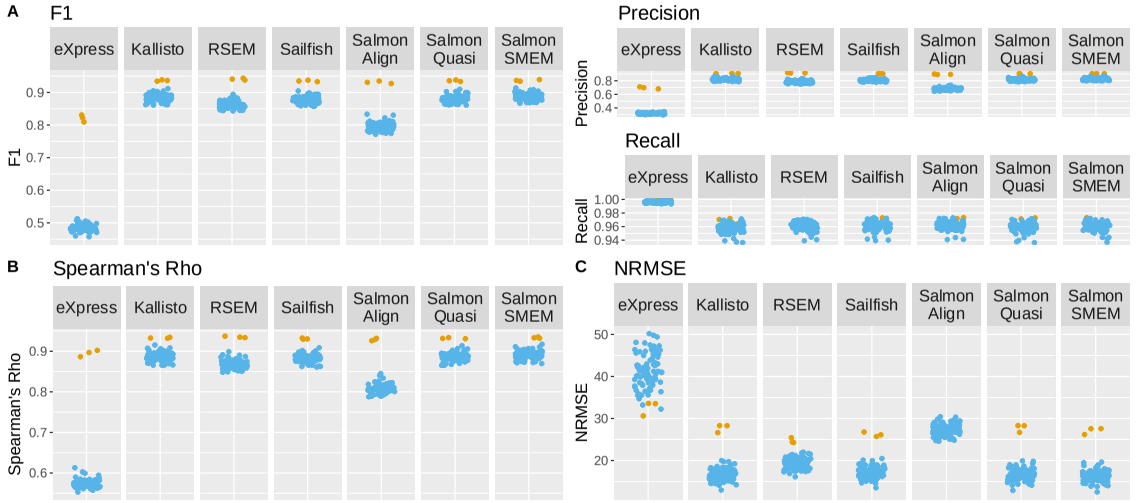


Figure 2.10: Comparison of the performance of isoform quantification tools on BLUEPRINT B lymphocyte bulk and single-cell RNA-seq data. Each point represents one cell from the scRNA-seq dataset or one bulk RNA-seq experiment. Yellow points represent bulk RNA-seq experiments, blue points represent one cell from the scRNA-seq experiment. **A** F1 score and precision and recall of isoform detection. The F1 score is the harmonic mean of the precision and recall. The precision is the proportion of the isoforms predicted to be expressed by an isoform quantification tool which are expressed. The recall is the proportion of expressed isoforms which are predicted to be expressed by the isoform quantification tool. **B** Spearman's rho. **C** Normalised root mean square error (NRMSE)

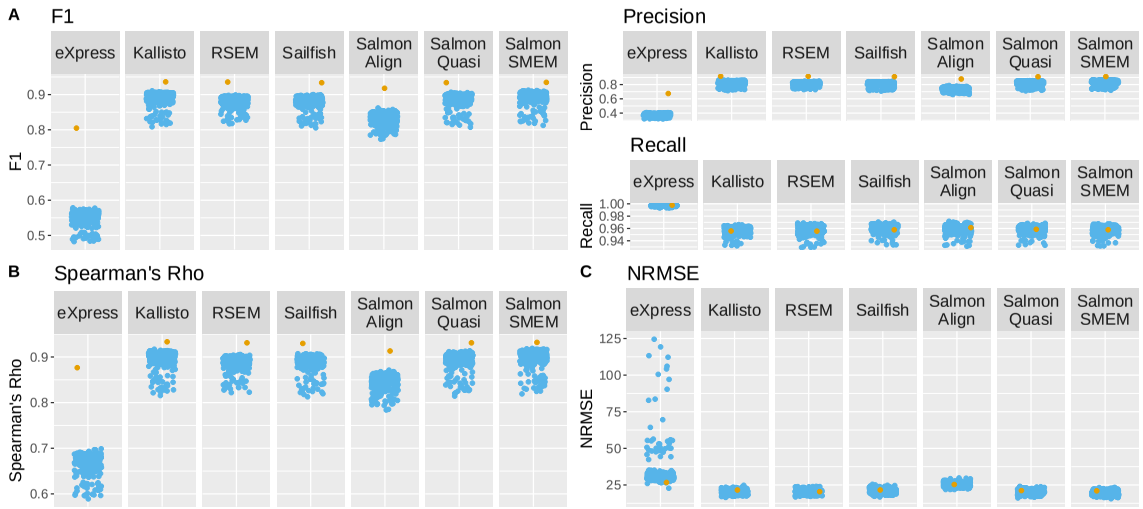


Figure 2.11: Comparison of the performance of isoform quantification tools on Kolodziejczyk et al. ES cell bulk and single cell RNA-seq data. Each point represents one cell from the scRNA-seq dataset or one bulk RNA-seq experiment. Yellow points represent bulk RNA-seq experiments, blue points represent one cell from the scRNA-seq experiment. **A** F1 score, precision and recall of isoform detection. The F1 score is the harmonic mean of the precision and recall. The precision is the proportion of the isoforms predicted to be expressed by an isoform quantification tool which are expressed. The recall is the proportion of expressed isoforms which are predicted to be expressed by the isoform quantification tool. **B** Spearman's rho. **C** Normalised Root Mean Square Error (NRMSE).

To gain further insights regarding the performance of the tools, I made use of the bulk RNA-seq data generated for the BLUEPRINT B lymphocytes and Kolodziejczyk et al. standard 2i media + LIF mESCs. I used RSEM to simulate the bulk RNA-seq data and collected the same performance statistics for my bulk RNA-seq benchmark as in my scRNA-seq benchmark. As the data used in my bulk and scRNA-seq benchmark came from the same source, the same method was used to generate the simulated bulk and scRNA-seq data, and the same performance statistics were collected in both benchmarks, I was able to carry out a meaningful comparison of isoform quantification tool performance on bulk and scRNA-seq data.

I find that all isoform quantification tools performed well on the simulated bulk data, but since most methods also performed well on single-cell data, the improvement was generally small (See Figure 2.10 and 2.11). In particular, there is very little difference in the recall for bulk and scRNA-seq, for which performance seems to be close to optimal. eXpress performs far better on bulk RNA-seq compared with scRNA-seq. Since eXpress appears to be overly liberal in calling isoforms as expressed, one explanation for the better performance of eXpress on bulk RNA-seq is that more isoforms have non-zero expression in bulk (see Appendix 1, Figure 7.11). Consequently, there are fewer unexpressed isoforms for eXpress to incorrectly call as expressed.

### 2.2.5 Removing drop-outs can improve the performance of isoform quantification tools.

While it is of interest to determine which isoform quantification tools perform best overall when run on scRNA-seq data, it is important to recognize that such an analysis may hide a lot of detail. For example, scRNA-seq data commonly contains a high number of dropouts (Ziegenhain et al., 2017), and one question of interest is whether the performance of isoform quantification tools differs between isoforms with a high number of dropouts and isoforms with few or no dropouts.

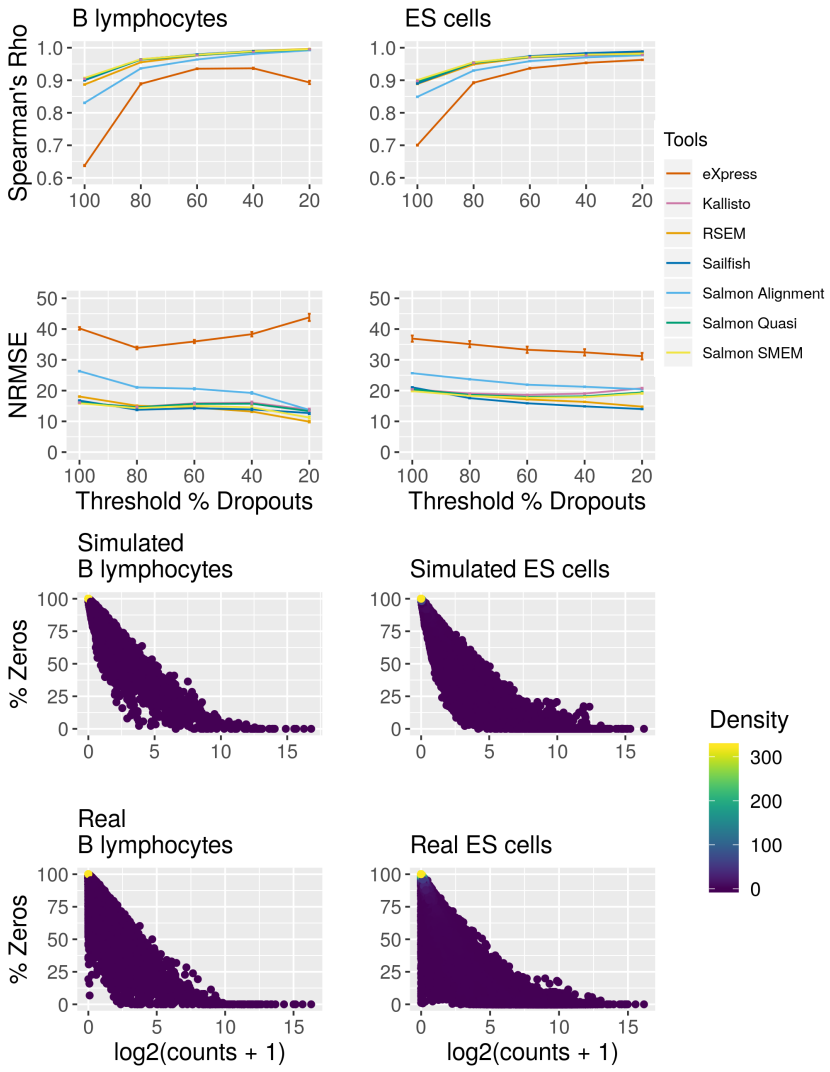


Figure 2.12: Effect of dropouts on isoform quantification. **A** Impact of removing isoforms with more than a threshold number of dropouts on Spearman's rho and the NRMSE for the BLUEPRINT B lymphocytes (left) and the Kolodziejczyk et al. ES cells (right). The x-axis gives the threshold percentage of zeros above which an isoform is removed from the analysis. For example, a threshold percentage of 80% would result in isoforms with zero expression in 80% or more of cells being removed from the analysis. Each colored line is a linear fit for visual guidance and it represents a different isoform quantification tool. **B** Relationship between how highly expressed an isoform is and the percentage of cells in which it has zero expression. The relationship is considered in RSEM simulated BLUEPRINT B lymphocytes (top left), the real BLUEPRINT B lymphocytes (bottom left), the simulated Kolodziejczyk et al. ES cells (top right), and the real Kolodziejczyk et al. ES cells (bottom right). Each point represents an isoform and the points are colored according to density.

To address the impact of dropouts on performance, Spearman's rho and the NRMSE were calculated when isoforms with zero expression in more than a specific fraction of cells were removed from the analysis. Applying increasingly stringent thresholds to remove isoforms with a high number of dropouts led to an increase in the value of Spearman's rho in both the Kolodziejczyk et al. and BLUEPRINT simulations (Figure 2.12A). For isoforms which had dropouts in less than 20% of cells, the value of Spearman's rho became very high for Sailfish, Salmon, Kallisto, and RSEM (in the range of 0.992–0.996 for the BLUEPRINT simulations, and 0.977–0.989 for the Kolodziejczyk et al. simulations). This indicates that for isoforms with very few dropouts, isoform quantification tools are extremely good at ordering their relative expression correctly. Removing isoforms with a high number of dropouts had a more variable effect on the NRMSE. Due to the inverse relationship between magnitude of expression and number of dropouts (Kharchenko et al., 2014) in both the real and simulated data (Figure 2.12B), one explanation for the increase in Spearman's rho is that lowly expressed isoforms are more likely to have a high number of dropouts and are also more likely to be mis-ordered with respect to the ground truth. However, because they are lowly expressed, removing them has a relatively small effect on the NRMSE.

### 2.2.6 The performance of Salmon alters depending on read depth.

So far in my benchmark, I have established that in general, isoform quantification tools perform well when run on full length scRNA-seq data with a moderately high read depth per cell and poorly when run on Drop-seq data with relatively low read depth per cell. To further investigate the impact of number of cells sequenced and number of reads sequenced per cell on the performance of isoform quantification tools, I return to the BLUEPRINT B lymphocyte dataset with a modified version of my simulation based approach. In my modified approach, I use RSEM to simulate data in which I systematically vary the number of cells simulated and the number of reads simulated per cell. The number of cells simulated was varied between 10

and 500 cells. As there are only 96 BLUEPRINT B lymphocytes, to generate my simulated cells I randomly selected cells from the BLUEPRINT B lymphocytes, then generated a simulated cell using RSEM run with a randomly selected seed. I varied the number of reads simulated from 0.25 million to 8 million reads per cell. For each datapoint, corresponding to a certain number of cells and a certain number of reads per cell, I performed 10 rounds of simulations. I then ran Salmon SMEM on each simulated cell and compared Salmon's expression estimates to the ground truth to generate performance statistics.

The reasons that Salmon and only Salmon was run on the simulated data are as follows. The total number of simulations performed in this experiment was vast (approximately 40,000), and consequently the computing resources required were also large, both in terms of time and memory (this experiment took over a month to run on the Sanger Institute's HPC). I have already established that the performances of Salmon, Sailfish, Kallisto and RSEM are very similar in multiple benchmarks, consequently the main goal of the experiment described here was to establish whether number of cells or number of reads per cell impact on isoform quantification tool performance in general, rather than which tool performed best at a given number of cells or read depth. This being so, in the interests of reducing computational time, I decided to benchmark a single tool rather than the full five. Salmon is one of the fastest and best performing tools investigated in this benchmark, consequently I chose to investigate Salmon's performance.

I first investigated how the number of cells simulated and the number of reads simulated per cell impacted on the number of expressed isoforms in my simulations and the number of isoforms detected by Salmon. The total number of expressed and detected isoforms increases as read number and cell number increases (Figure 2.13). However, the total number of detected isoforms vastly exceeds the number of expressed isoforms – over 6000 isoforms were detected when 500 cells were simulated with 8 million reads per cell, but less than 4000 isoforms were expressed in the ground truth. This means over one third of the total number of detected isoforms were false positives, a finding with important implications for studies attempting to determine the number of expressed genes and isoforms in a population of cells.

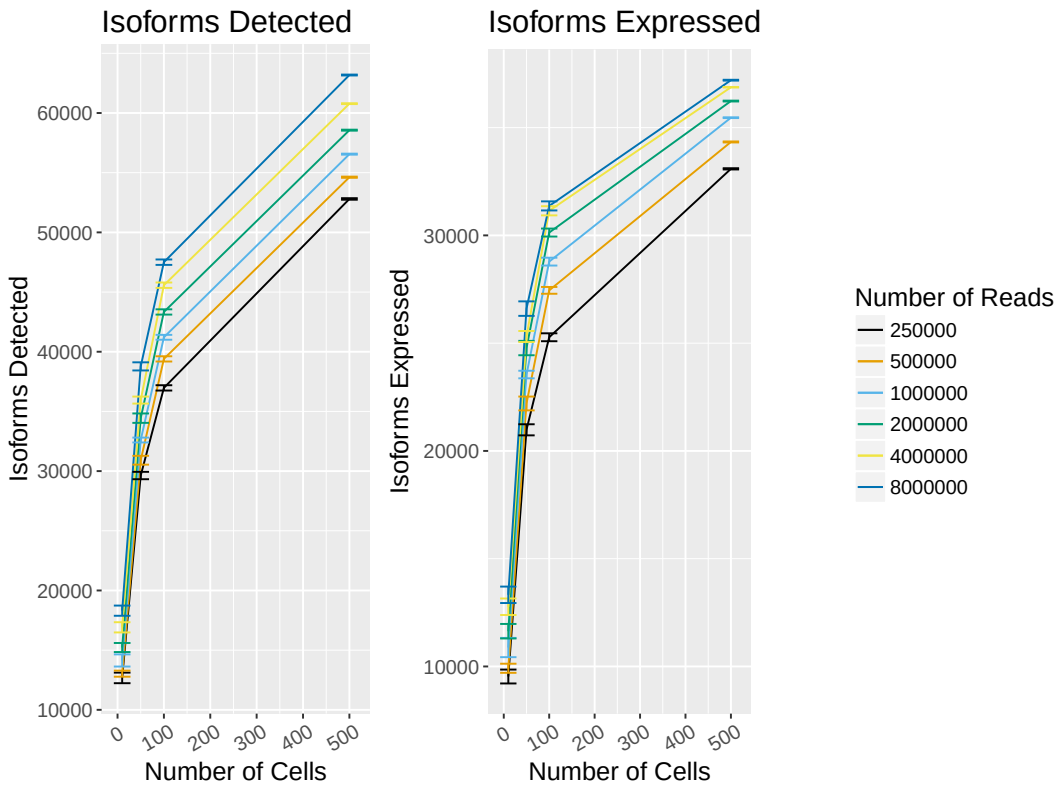


Figure 2.13: The difference between the total number of isoforms detected as expressed by Salmon SMEM (left) and the number of isoforms expressed in the ground truth (right) in RSEM simulated data. Each coloured line is a linear fit for visual guidance and represents the number of reads sequenced per cell.

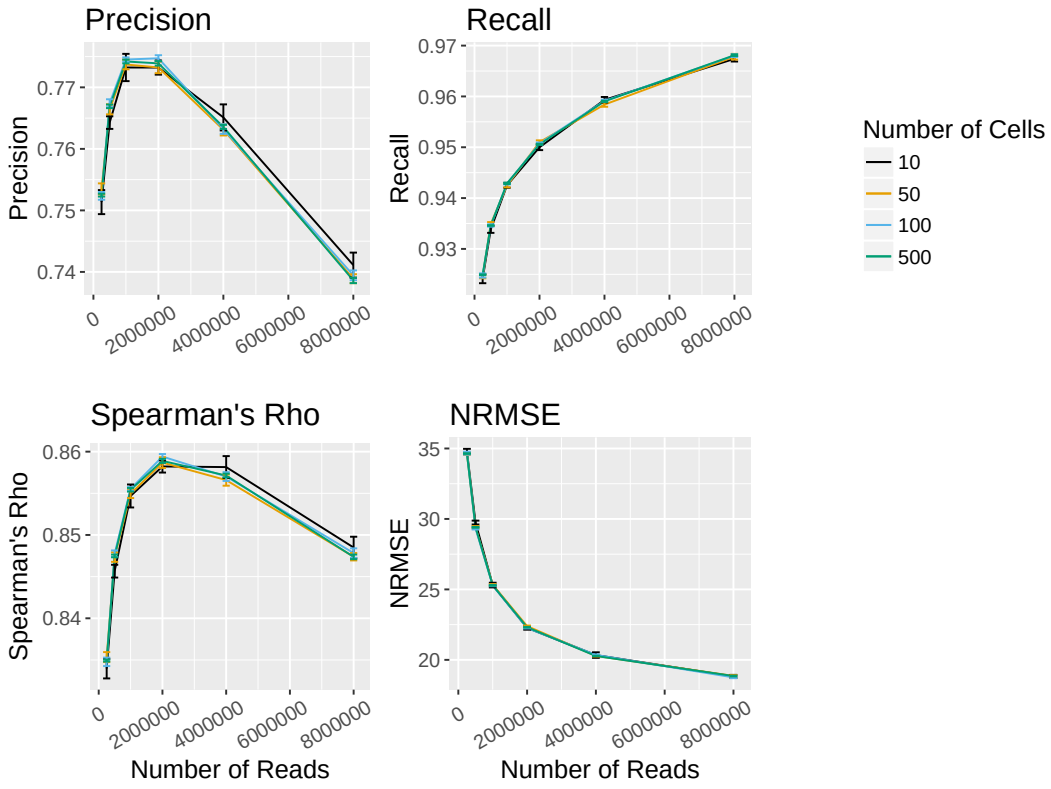


Figure 2.14: The impact of the number of cells sequenced and the number of reads sequenced per cell on the performance of Salmon run on SMEM mode. Each coloured line is a linear fit for visual guidance and represents the number of cells sequenced.

I next considered the impact of the number of cells and the number of reads per cell on performance. The number of cells simulated had little impact on performance, however the number of reads simulated per cell had an impact on Spearman's Rho, the NRMSE and the precision and recall of isoform detection (Figure 2.14). As read number increased, the recall of isoform detection increased and the NRMSE decreased. It is perhaps unsurprising that the ability of Salmon to detect and accurately quantify expressed isoforms increases as the number of reads increases. More surprisingly, the precision of isoform detection and Spearman's Rho peak at around 2 million reads then decrease as the number of reads further increase. To investigate this result further, I decided to focus on the precision of isoform detection. The precision is defined as:

$$Precision = \frac{NumberOfTruePositives}{NumberOfTruePositives + NumberOfFalsePositives}$$

Where I define a true positive as an isoform which is called as expressed by the isoform quantification tool which is expressed in the ground truth, and a false positive as an isoform which is called as expressed by the isoform quantification tool which is not expressed in the ground truth. As the precision depends on the number of true positives and false positives, I investigated the relationship between the number of true and false positives and the read number. I found that the number of true positives plateaus over the range of read numbers considered in my experiment whereas the number of false positives does not (Figure 2.15). Thus, the peak in the precision is a consequence of the number of false positives increasing more rapidly than the number of true positives at high read numbers. The slower rate of increase in the number of true positives with increasing read numbers is most likely a consequence of the plateau in the number of expressed isoforms at high read numbers in the simulated data (Figure 2.16). Note that RSEM generates ground truth expression measures by counting where each of the reads it simulates originated in the transcriptome, thus unlike in real scRNA-seq data, there are no expressed isoforms from which no reads are captured. Nonetheless, it is true that a sequenced cell must express a fixed number of isoforms and that the number of true positives

will be limited at high read numbers by the number of isoforms it expresses. My findings suggest there is an optimum number of reads to sequence per cell if the aim of the experiment is isoform detection, and that the optimum number may be partly determined by the number of isoforms expressed by the cell.

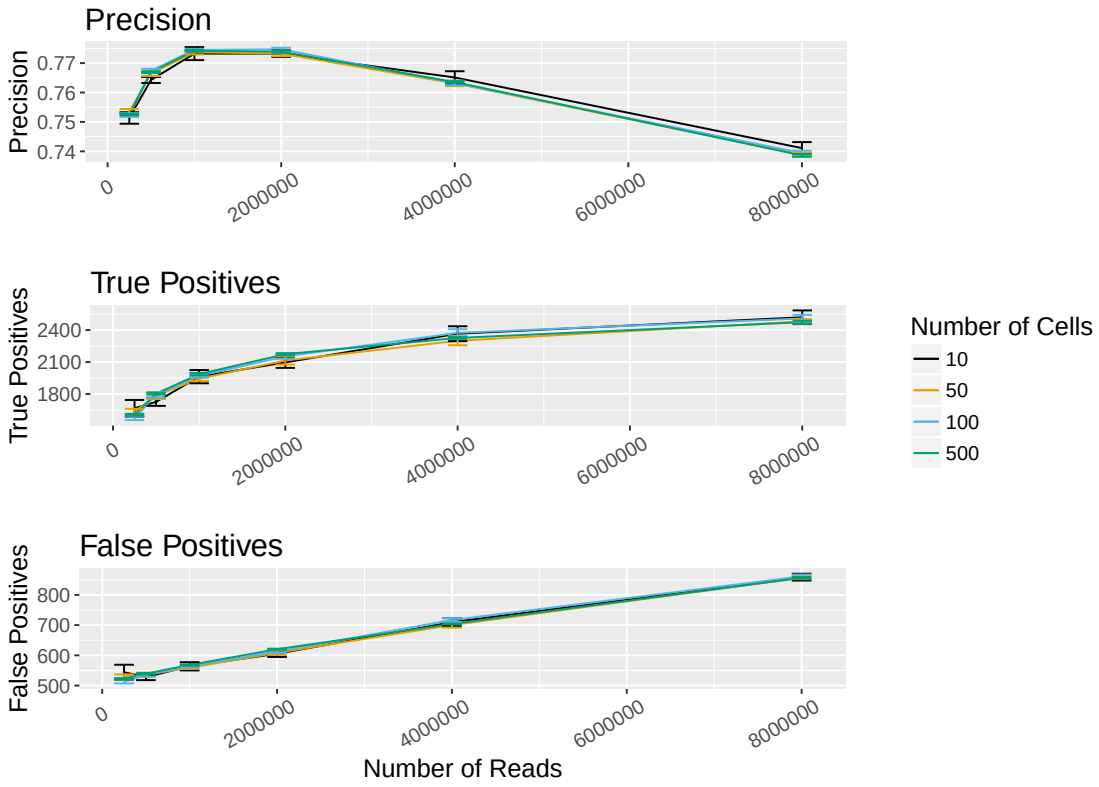


Figure 2.15: The impact of the number of reads sequenced per cell on the precision, the number of true positives and the number of false positives. In this context, a true positive is an isoform called as expressed by Salmon SMEM which is expressed in the ground truth. A false positive is an isoform called as expressed by Salmon SMEM which is unexpressed in the ground truth. Each coloured line is a linear fit for visual guidance and represents the number of cells sequenced.

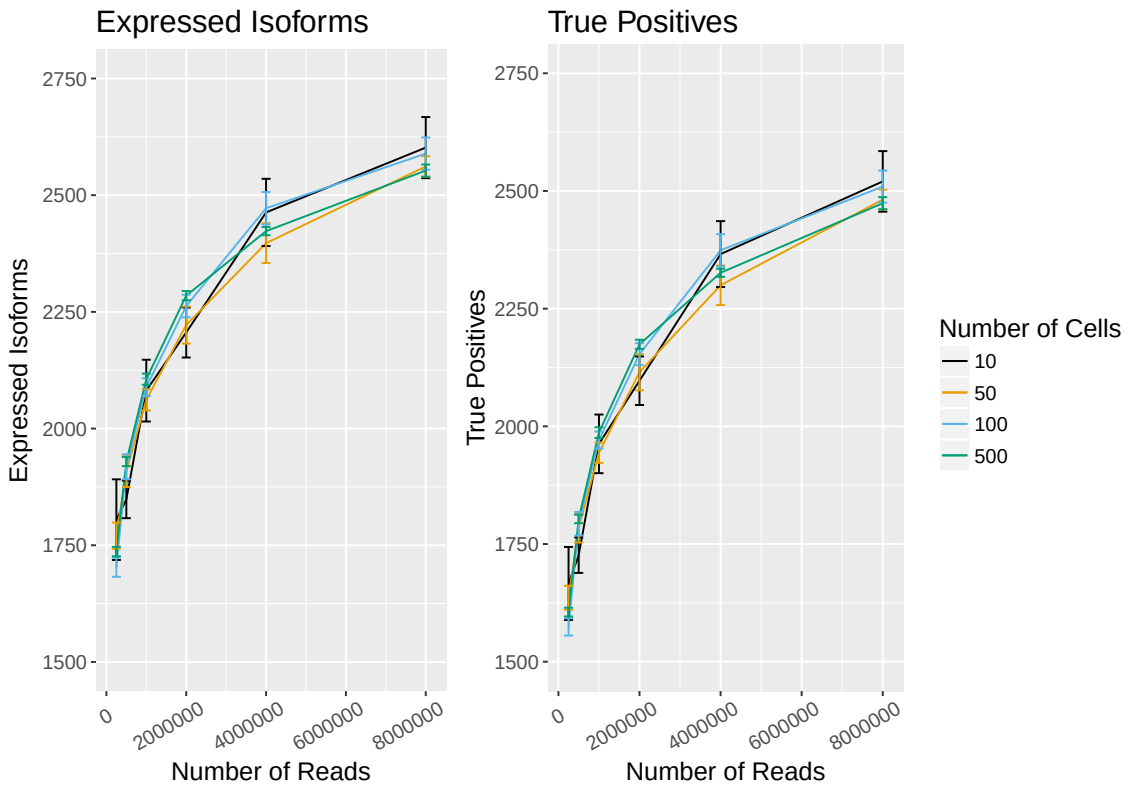


Figure 2.16: The effect of read number on the number of expressed isoforms (left) and the number of true positives (right). In this context, a true positive is an isoform called as expressed by Salmon SMEM which is expressed in the ground truth. Each coloured line is a linear fit for visual guidance and represents the number of cells sequenced.

## 2.3 Discussion

To date, scRNA-seq studies have mainly focussed on gene level quantification (Stegle et al., 2015). This has partly been due to uncertainty over how best to perform isoform quantification in scRNA-seq. In addition, there has been uncertainty over whether the results obtained would be meaningful due to the low read coverage compared with bulk RNA-seq. My analyses have demonstrated that Kallisto, Salmon, Sailfish and RSEM can accurately detect and quantify isoforms in scRNA-seq to nearly the same accuracy as bulk RNA-seq data, provided the datasets have a reasonably high number of reads per cell. For simulated data based on Drop-seq, the performance of isoform quantification tools was too poor to make the results of performing isoform quantification worthwhile. Due to the low precision of isoform detection, this problem cannot be overcome by incorporating information from a large number of cells. It is possible that increasing the number of reads sequenced per cell for Drop-seq protocols would improve the performance of isoform quantification tools, although the short reads and 3' coverage bias are likely to ensure that accurate quantification remains challenging.

A potential limitation of my benchmark is that we might expect that the degree of cellular heterogeneity might differ between quiescent B lymphocytes, mESCs and retinal bipolar cells, potentially confounding the comparison of the three sequencing technologies. As the benchmark is based on a comparison within each simulated cell between its ground truth expression and quantification tool expression estimates, I would expect structural differences in the real datasets to be a relatively minor confounder. The confounder could be removed by repeating the benchmark with a cell population that was sequenced using several different sequencing technologies.

A systematic investigation found that the number of cells sequenced has no perceptible impact on the performance of isoform quantification. I find that the precision of isoform detection peaked in my simulations at around 1-2 million reads per cell. I hypothesise that this could occur because that the number of expressed isoforms in my simulated data does not increase substantially beyond 2 million reads per cell, and the majority of expressed isoforms are already called as expressed at 2 mil-

lion reads. Consequently, beyond 2 million reads per cell, it is not possible for the number of true positives to increase much further, whereas the number of false positives continues to increase. The position of the peak at 1-2 million reads per cell is possibly an RSEM simulation artefact, as it likely occurs due to RSEM not substantially increasing the number of isoforms expressed per gene per cell beyond 2 million reads per cell. However, it is a fact cells express a finite number of isoforms, thus I would predict that the precision of isoform detection will also peak at a particular read depth in real scRNA-seq data. The position of the peak would depend on the number of isoforms expressed by each cell, so consequently might vary between cell types and species. This observation is highly relevant when analysing very deeply sequenced scRNA-seq data, as it predicts that an increasing proportion of detected isoforms and genes will be false positives at very high read depths.

In addition to benchmarking isoform quantification for scRNA-seq, I perform an equivalent benchmark for bulk RNA-seq. I find that the performance of most isoform quantification tools is slightly worse for scRNA-seq compared with bulk, but that the difference is small. The cost in performance using scRNA-seq compared with bulk RNA-seq for isoform quantification is therefore low. However, it should be noted that this benchmark has evaluated the ability of isoform quantification tools to correctly assign the reads present in an scRNA-seq experiment to the isoforms they originated from. As a major technical issue with scRNA-seq is failure to capture reads from a high proportion of expressed transcripts (Vallejos et al., 2017), it is likely that in practice, many expressed isoforms will be missed by isoform quantification tools when run on scRNA-seq data due to a lack of evidence in the captured reads that the isoform was expressed. However, the extremely high recall of all the isoform quantification tools considered in this benchmark means that the overwhelming majority of isoforms from which reads are captured will be called as expressed. More problematic is the relatively low precision of isoform detection, as a consequence of which around one in six isoforms called as expressed in deeply sequenced scRNA-seq datasets will be false positives, even for the best performing tools.

Whilst my analysis has demonstrated that existing tools can accurately detect and quantify isoforms for scRNA-seq, no tool performed perfectly. The tools bench-

marked here were designed for use with bulk RNA-seq, and it is plausible that future tools designed to perform isoform quantification specifically for scRNA-seq could perform better. I found that the tools benchmarked in this study tended to have a higher recall than precision of isoform detection. Therefore, it is likely the performance of isoform quantification tools designed for scRNA-seq data could be improved by making the tools more conservative in calling isoforms as expressed relative to tools designed for use on bulk data. In addition, I found that Spearman's rho increased when lowly expressed isoforms with a high number of dropouts were removed from the analysis. Thus, it is likely that attempts to incorporate the effects of single cell specific technical noise such as dropouts would improve the performance of isoform quantification tools on scRNA-seq. An open question for isoform quantification in scRNA-seq is whether incorporating information from Unique Molecular Identifiers (UMIs) into isoform expression estimates could improve accuracy of quantification. Whilst UMI information could reduce the effects of PCR amplification noise (Islam et al., 2014), UMI based protocols tend to exhibit significant coverage bias, potentially making isoform quantification challenging (Grün and van Oudenaarden, 2015). If UMI based protocols could be combined with long read sequencing technologies, this problem could potentially be overcome.

## 2.4 Conclusions

For high-quality simulated scRNA-seq datasets with a high number of reads/cell, RSEM, Kallisto, Salmon, and Sailfish can accurately detect and quantify isoform expression. Isoforms with a high number of dropouts appear to be relatively challenging to quantify, possibly because such isoforms are often lowly expressed. In my benchmark of bulk RNA-seq, I discover the performance of most isoform quantification tools is slightly worse for scRNA-seq compared with bulk, but that the difference is small.

Taken together, my findings show that isoform quantification is possible with scRNA-seq for SMARTer and SMART-seq2 data. As single cells do not generally express all of the isoforms seen at the population level, scRNA-seq may eventu-

ally provide advantages over bulk RNA-seq for isoform quantification by essentially deconvoluting the problem of isoform quantification. Future isoform quantification tools designed explicitly for scRNA-seq could improve on the performance of existing tools by being more conservative in calling isoforms as expressed, and by explicitly modeling the technical noise inherent to scRNA-seq.

# 3

# Attempts to Determine How Many Isoforms Are Produced per Gene per Cell Give Uninterpretable Results.

*Science, my boy, is made up of mistakes, but they are mistakes which it is useful to make, because they lead little by little to the truth.*

– Jules Verne, *Journey to the Centre of the Earth*(Verne, 1864)

## 3.1 Introduction

In the previous chapter, I demonstrated that Kallisto, Salmon, Sailfish and RSEM perform almost as well when run on scRNA-seq as when run on bulk RNA-seq. This is an exciting result, because it suggests that existing isoform quantification software performs sufficiently well on scRNA-seq to theoretically enable alternative splicing to be studied using scRNA-seq data. In this chapter, I therefore design a series of experiments to answer a basic biological question related to splicing using scRNA-seq.

'How many isoforms does a gene produce per cell' is a fundamental question in molecular biology, yet for most genes we do not have an answer to this question. In addition to being of interest to basic biologists, establishing how isoform expression is regulated at a cellular level could potentially be of therapeutic relevance to the many patients suffering from diseases in which splicing has been implicated. If insight into splicing regulation at a cellular level could be gained from scRNA-seq data, a far higher number of genes could be studied at much lower cost than would be possible using lower throughput approaches such as smFISH. In an attempt to shed some light on how isoform expression is regulated in cells, I designed a series of experiments investigating isoform number in individual cells using scRNA-seq. Unfortunately, my main conclusion from these experiments was that without a better understanding of the technical noise associated with scRNA-seq, it is often not possible to distinguish between genuine splicing behaviour and technical noise in individual cells.

Some of the work presented in this chapter has been published, consequently some passages have been quoted verbatim from the following sources: (Westoby et al., 2018a,b, 2019). Additionally, some figures have been reproduced from the aforementioned sources.

## 3.2 Results

### 3.2.1 For genes which express two isoforms in bulk RNA-seq, usually only one isoform is detected per cell in scRNA-seq.

To determine whether individual cells express all or only some of the isoforms seen in a population of cells, I consider genes which have two isoforms, both of which are expressed in the BLUEPRINT B lymphocyte or in the Kolodziejczyk et al. ES cell bulk RNA-seq data. I then determine how many isoforms are expressed from these genes in the corresponding scRNA-seq data. Kallisto was used to perform isoform quantification for the bulk and single-cell data as it performed well in both my bulk

RNA-seq and scRNA-seq benchmarks.

For genes which express two isoforms in bulk RNA-seq data, I first consider if zero, one, or two isoforms are detected in single cells. For most genes which express two isoforms in the bulk RNA-seq, neither isoform is detected in most cells in the scRNA-seq (Figure 3.1A & 3.2A). A biological interpretation could be that the gene expression detected in bulk RNA-seq reflects heterogeneous gene expression at a cellular level. However, an arguably more realistic technical interpretation is that technical dropouts are prevalent in scRNA-seq (Kharchenko et al., 2014; Marinov et al., 2014; Svensson et al., 2017; Islam et al., 2014) and consequently I fail to detect expressed genes in many cells. Of course, the two explanations are not necessarily exclusive. It is possible that there is both heterogeneous gene expression, and that I fail to detect gene expression in many cells due to dropouts.

In cells where gene expression is detected, it is more common to detect one rather than two isoforms. To investigate this further, I consider the percentage of cells in which both of the isoforms expressed in the bulk RNA-seq data are detected. I find that for the majority of genes, both isoforms are detected in no or very few cells; however, for a minority of genes in both the BLUEPRINT B lymphocytes and Kolodziejczyk et al. ES cells, both isoforms are detected in a high percentage of cells (Figure 3.1B & 3.2B). There are more genes for which both isoforms are detected in the Kolodziejczyk et al. ES cells compared to the BLUEPRINT B lymphocytes. This may partly reflect the higher number of cells and the higher number of reads per cell in the Kolodziejczyk et al. ES cells, possibly enabling better detection of lowly and/or infrequently expressed isoforms. In addition, the globally elevated transcription rates in ES cells relative to other cell types might lead us to expect that expression of multiple isoforms from a single gene would be more common in ES cells (Efroni et al., 2008), especially compared to quiescent and transcriptionally inactive B lymphocytes.

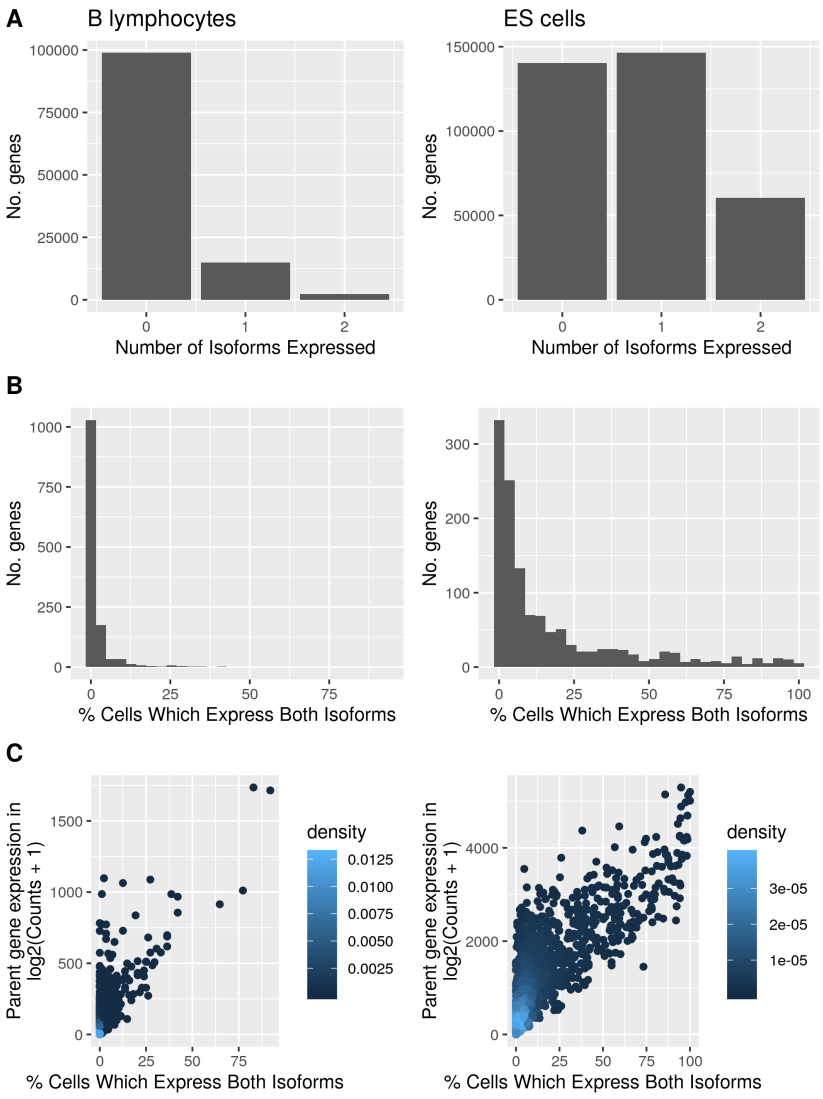


Figure 3.1: Investigation into how many isoforms are expressed per cell in the scRNA-seq data for genes which express exactly two isoforms in bulk data. The BLUEPRINT B lymphocyte (left) and the Kolodziejczyk et al. ES cell (right) bulk RNA-seq data are shown. The B lymphocyte graphs shown here are from the first biological replicate of the BLUEPRINT male B lymphocyte bulk RNA-seq; equivalent graphs for the second and third BLUEPRINT male B lymphocyte biological replicates can be found in Figure 3.2. **A** Number of genes which express two isoforms in the bulk RNA-seq data expressing zero, one or two isoforms in each cell in the scRNA-seq data. **B** Histogram of the percentage of cells which express both the isoforms detected in the bulk RNA-seq data. **C** Relationship between the percentage of cells which express both the isoforms detected in the bulk RNA-seq. The y axis represents the log transformed parent gene expression, found by summing the expression of the two isoforms. Spearman's rho is 0.623 for the BLUEPRINT B lymphocytes and 0.795 for the Kolodziejczyk et al. ES cells. Points are coloured by density. Note that the B lymphocyte replicates appear extremely similar because the replicates refer to bulk RNA-seq replicate samples. The same scRNA-seq B lymphocyte dataset is used each time.

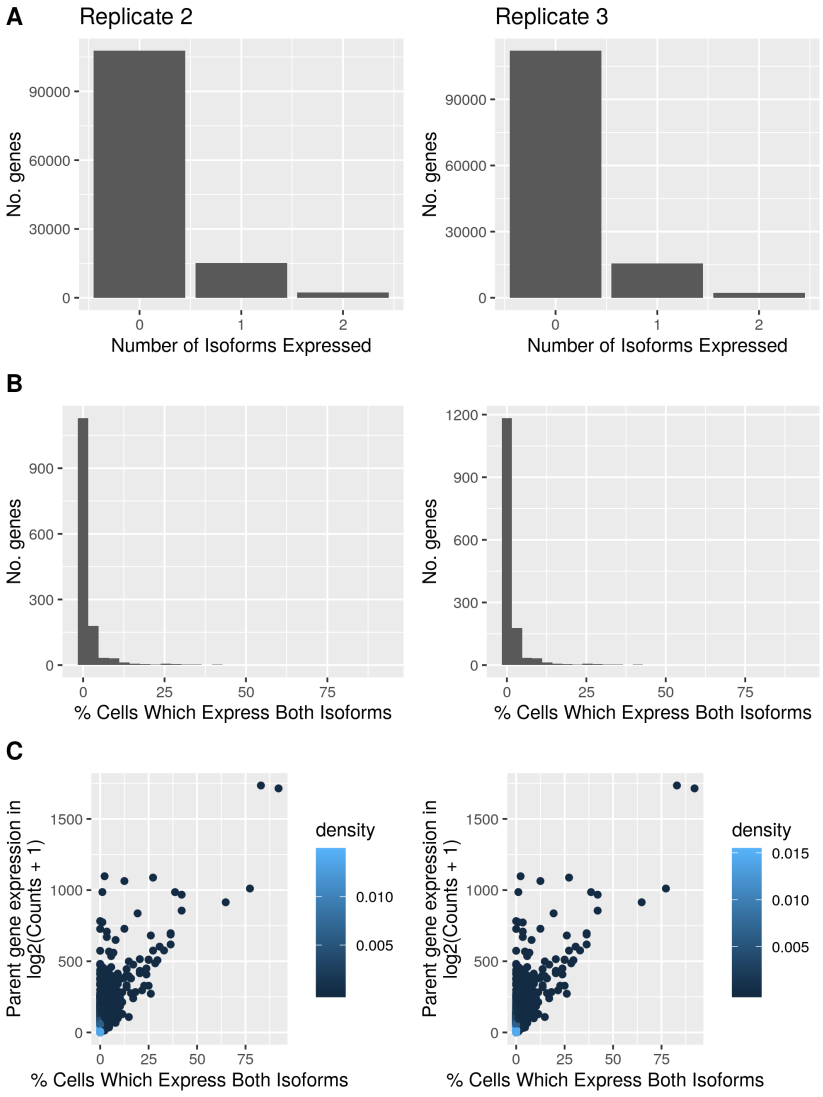


Figure 3.2: Investigation into how many isoforms are expressed per cell in the scRNA-seq data for genes which express exactly two isoforms using the second (left) and third (right) biological replicates for the BLUEPRINT B lymphocyte bulk RNA-seq data. A: Number of genes which express two isoforms in the bulk RNA-seq data which express zero, one or two isoforms in each cell in the scRNA-seq data. B: Histogram of the percentage of cells which express both the isoforms detected in the bulk RNA-seq data. C: Relationship between the percentage of cells which express both the isoforms detected in the bulk RNA-seq. The y axis represents the log transformed parent gene expression, found by summing the expression of the two isoforms. Spearman's rho is 0.625 for replicate 2 and 0.626 for replicate 3. Points coloured by density. Note that the B lymphocyte replicates appear extremely similar because the replicates refer to bulk RNA-seq replicate samples. The same scRNA-seq B lymphocyte dataset is used each time.

Finally, I ask whether more highly expressed genes are more likely to express multiple isoforms. I find a positive correlation between gene expression and the percentage of cells in which both isoforms are detected (Figure 3.1C & 3.2C). The observation that in scRNA-seq, it is more common to detect multiple isoforms in individual cells for highly expressed genes relative to lowly expressed genes is not new (Zhao et al., 2016; Marinov et al., 2014). Marinov et al. proposed that this reflects a biological phenomenon and that more highly expressed genes on average undergo more alternative splicing and produce more isoforms in individuals cells compared to lowly expressed genes (Marinov et al., 2014). However, there is also a potential technical explanation. The probability of a transcript becoming a dropout is inversely proportional to how highly expressed that transcript is (Kharchenko et al., 2014). Consequently, if isoform expression was identical in all cells, we would expect to be able to detect highly expressed isoforms in a higher proportion of cells and lowly expressed isoforms in a lower proportion of cells. It is likely that many isoforms from highly expressed genes are themselves highly expressed. Therefore, one explanation for the correlation between magnitude of gene expression and percentage of cells in which both isoforms are detected is that the probability of detecting both isoforms increases as gene expression increases. In other words, it is possible that lowly expressed genes produce both isoforms in individual cells, but the probability of dropout is so high that we fail to detect them.

### 3.2.2 A novel simulation approach suggests that Tbx3, Klf4 and Pou5f1 are differentially spliced in mESCs cultured in different conditions.

The results shown in Figure 3.1 & 3.2 indicate that for genes where two isoforms are detected in bulk RNA-seq, it is rare to detect both isoforms in scRNA-seq, but more common for highly expressed genes. Without knowing how to appropriately correct for dropouts, it is challenging if not impossible to distinguish whether this reflects biological reality or technical noise.

I rationalised that a different approach, in which dropouts were explicitly ac-

counted for, was required to analyse alternative splicing using scRNA-seq. I therefore designed a novel simulation based approach. In my approach, a gene of interest is selected and the total number of detected isoforms over an entire real scRNA-seq dataset, $N$, is established. I then simulate $N$ scenarios. In the first scenario, I simulate a situation in which 1 isoform is expressed from the gene of interest in all cells. In the second scenario, I simulate a situation in which 2 isoforms are expressed in all cells, and so on up to the $N$th scenario, where $N$ isoforms are expressed in all cells. Importantly, in each scenario, I simulate dropout events using a Michaelis-Menten model developed by Andrews and Hemberg (Andrews and Hemberg, 2018a), and simulate quantification errors using error rates estimated from my benchmark in chapter 2. I record the number of detected isoforms in each cell after dropouts and quantification errors are simulated, then find the mean number of isoforms detected across all cells. I then repeat this process thousands of times to generate distributions of the mean number of isoforms detected per cell when $n$ in range 1:$N$ isoforms are expressed per cell. This process is illustrated in Figures 3.3 and 3.4.

The distributions of the mean number of isoforms detected per cell generated by my simulations can be considered to be null distributions - they are the distributions of the mean number of isoforms detected per cell if exactly $n$ isoforms are expressed in every cell, assuming that I am modelling dropouts, isoform choice and quantification errors appropriately. Onto these distributions, I draw a vertical black line representing the mean number of isoforms detected per cell in the real data. If the black line falls to the right of the distribution for $n$ isoforms, or into the 2.5% largest values of the distribution, my simulation model predicts that cells produce significantly more than $n$ isoforms per cell in reality. If the black line is to the left of the $n$ isoform distribution, or in the 2.5% lowest values of the distribution, my model predicts that cells produce significantly less than $n$ isoforms per cell. If the black line falls into the middle 95% of values of the $n$ isoform distribution, my model predicts that cells produce $n$ isoforms per cell.

To test the predictive ability of my model, I returned to the Kolodziejczyk et al. ES cell dataset. The mESCs in this dataset were cultured under three different cuture conditions (Kolodziejczyk et al., 2015). One culture condition consisted of

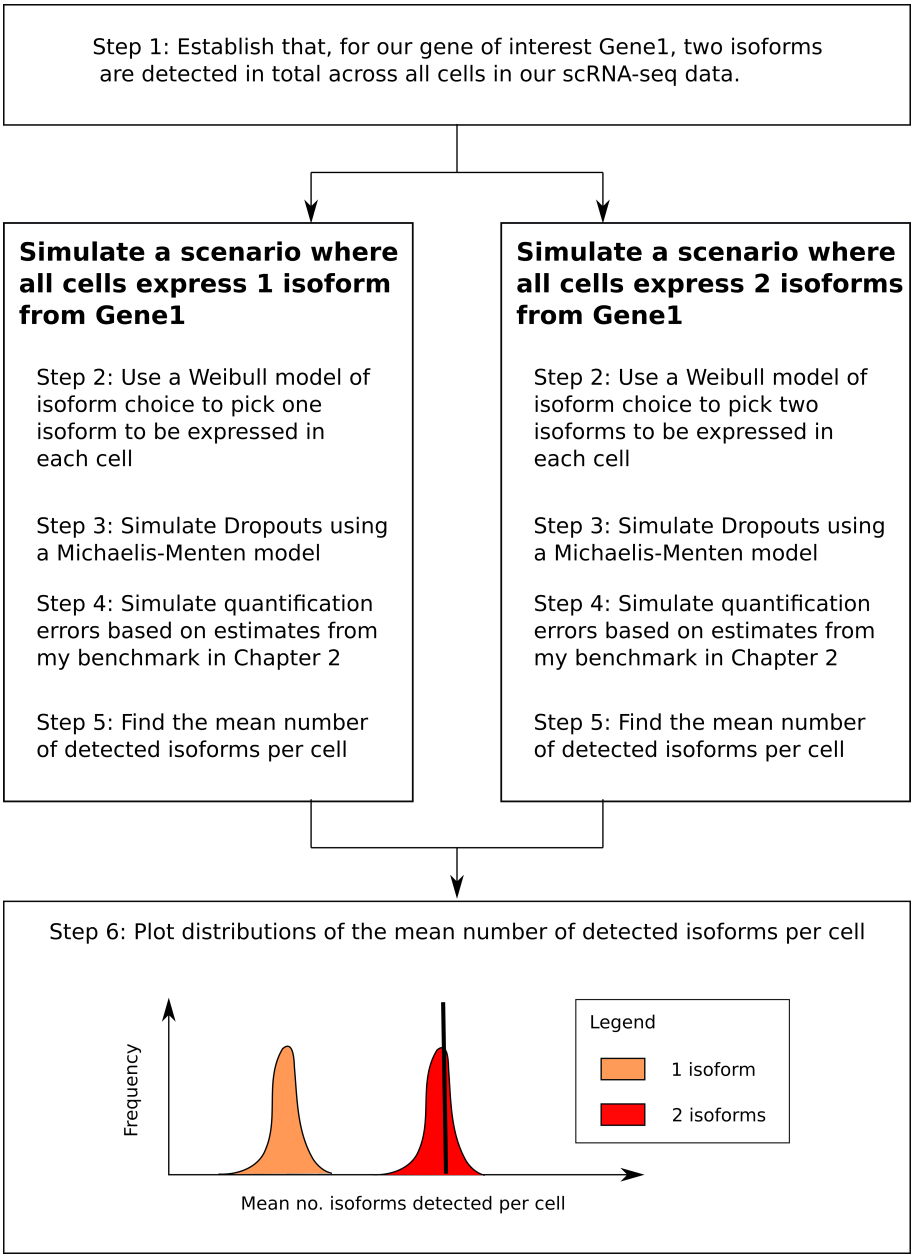


Figure 3.3: Schematic of my simulation approach. The final output is a graph showing $N$ distributions, each corresponding to the mean number of isoforms detected when $n$ isoforms are expressed per gene per cell. The black line shown on the graph is the mean number of isoforms detected per gene per cell in the real data - in this case, the mean number of isoforms detected in the real data is consistent with two isoforms being expressed in every cell.

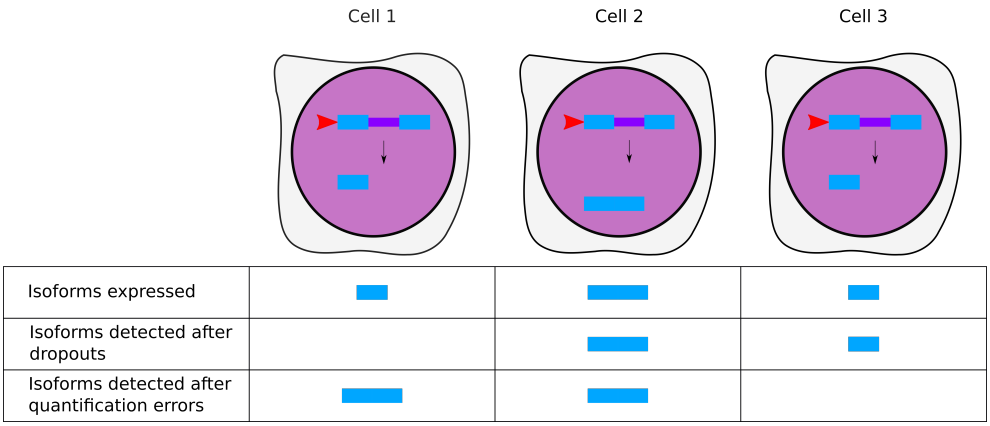


Figure 3.4: An example of the output at each stage of my simulation approach. For example, simulated Cell 1 expresses the short isoform from our gene of interest. However, the short isoform is not detected due to a dropout event. A quantification error then occurs, leading to the long isoform being detected, despite the long isoform never having been expressed in Cell 1.

serum + LIF, which I will refer to as the 'serum' culture condition, one culture condition was 2i + LIF, which I will refer to as 'standard 2i', and one culture condition was a2i + LIF, which I will refer to as 'a2i'. It is known that the morphology and transcriptional properties of mESCs can substantially differ depending on the conditions that they are cultured in (Morgani et al., 2017; Marks et al., 2012). In addition, it has been observed that the expression of some pluripotency factors is more heterogeneous in mESCs cultured in serum/LIF than in mESCs cultured in 2i (Toyooka et al., 2008; Chambers et al., 2007). It would be interesting to establish whether key players in the pluripotency network are differentially spliced when mESCs are cultured in different conditions. If differential splicing is occurring, it could partly explain the morphological and transcriptional differences that have been observed between culture conditions.

To investigate whether key players in the pluripotency network are differentially spliced between mESC culture conditions, I applied my model to 15 genes implicated in playing a role in maintaining pluripotency (see Methods chapter for a complete list). These 15 genes were chosen as a starting point to this study due to their well

established role in the pluripotency network. Of these 15 genes, my model suggested that 7 genes (Klf4, Pou5f1, Tbx3, Jarid2, Myc, Stat3 and Tcf3) produced differing numbers of isoforms depending on culture condition (Figures 3.5-3.11). Of these 7 genes, 4 genes (Jarid2, Myc, Stat3, Tcf3) had a very large number of detected isoforms (Figures 3.8-3.11). I would expect isoform quantification to be more error prone when genes produce a very large number of similar isoforms, and therefore have less confidence in my model's predictions for these genes. Consequently, I decided to focus on the 3 remaining genes (Tbx3, Klf4 and Pou5f1) for which six or less isoforms were detected as candidates for differential splicing between culture conditions.

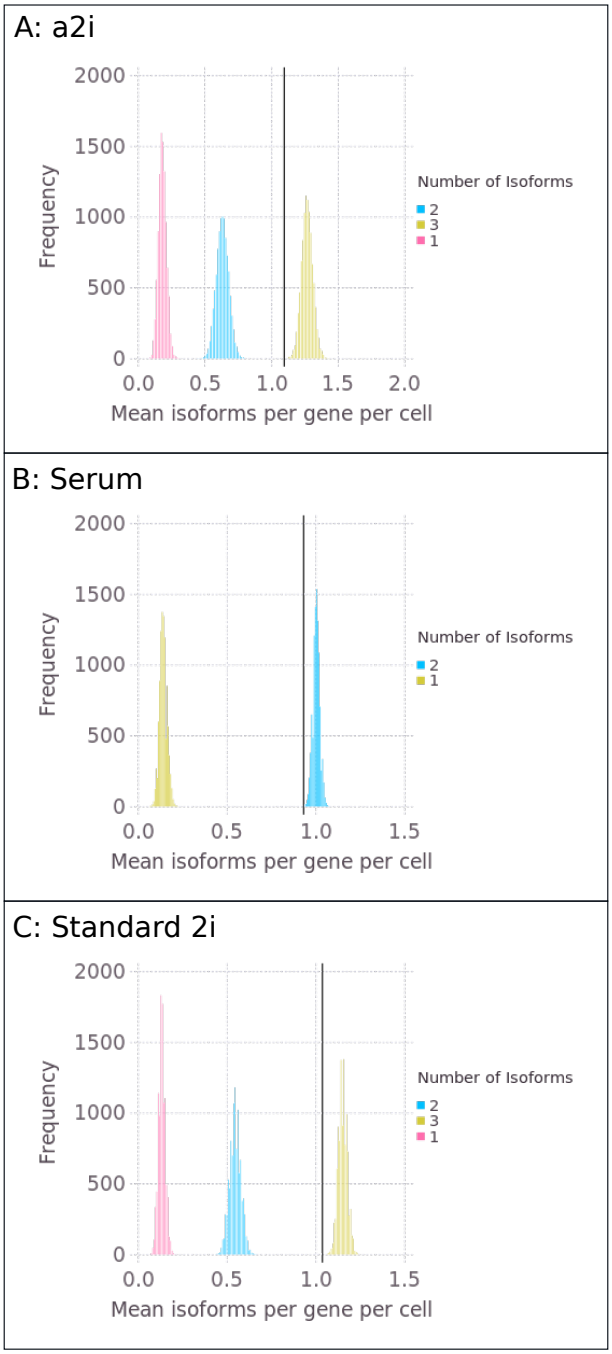


Figure 3.5: Simulation results for Klf4 gene in mESCs cultured in **A** a2i culture conditions, **B** serum and **C** standard 2i culture conditions. The vertical black line on each plot represents the mean number of isoforms detected per cell in the real data.

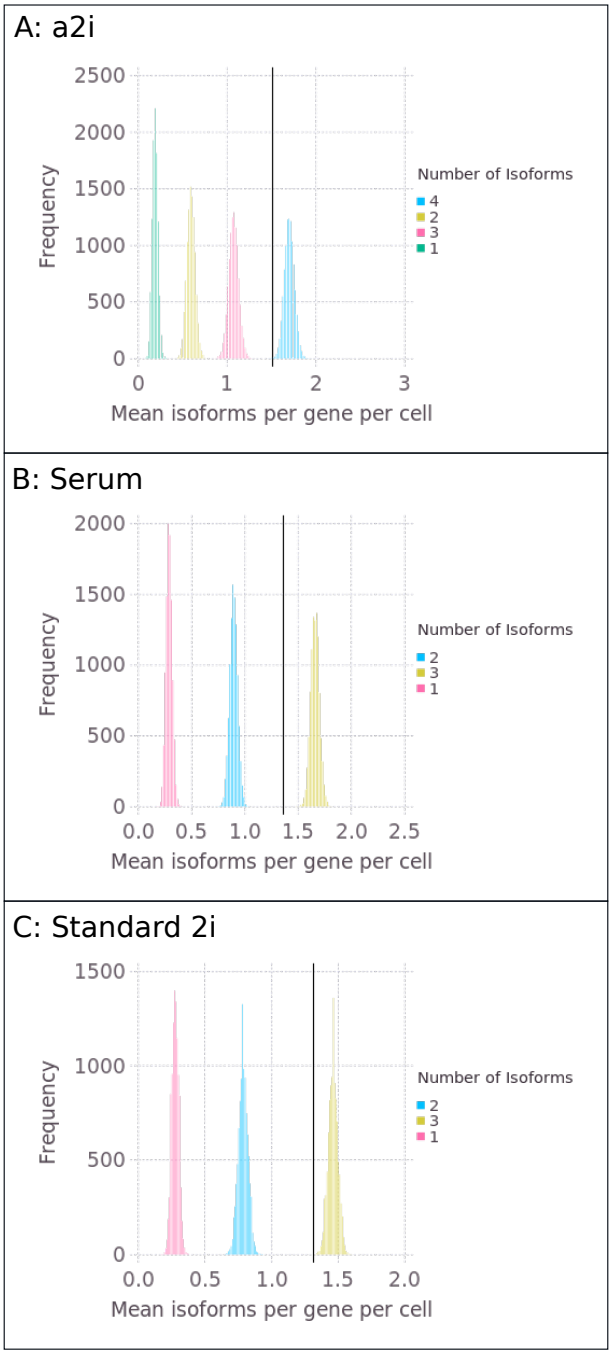


Figure 3.6: Simulation results for Pou5f1 gene in mESCs cultured in **A** a2i culture conditions, **B** serum and **C** standard 2i culture conditions. The vertical black line on each plot represents the mean number of isoforms detected per cell in the real data.

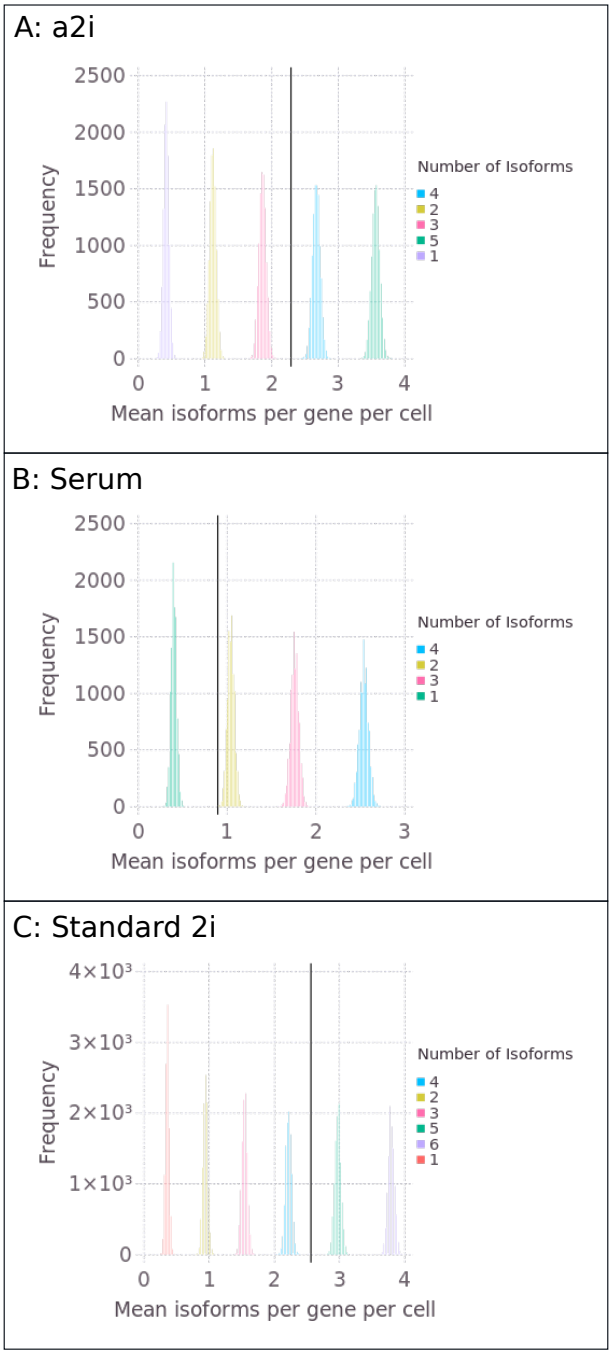


Figure 3.7: Simulation results for Tbx3 gene in mESCs cultured in **A** a2i culture conditions, **B** serum and **C** standard 2i culture conditions. The vertical black line on each plot represents the mean number of isoforms detected per cell in the real data.

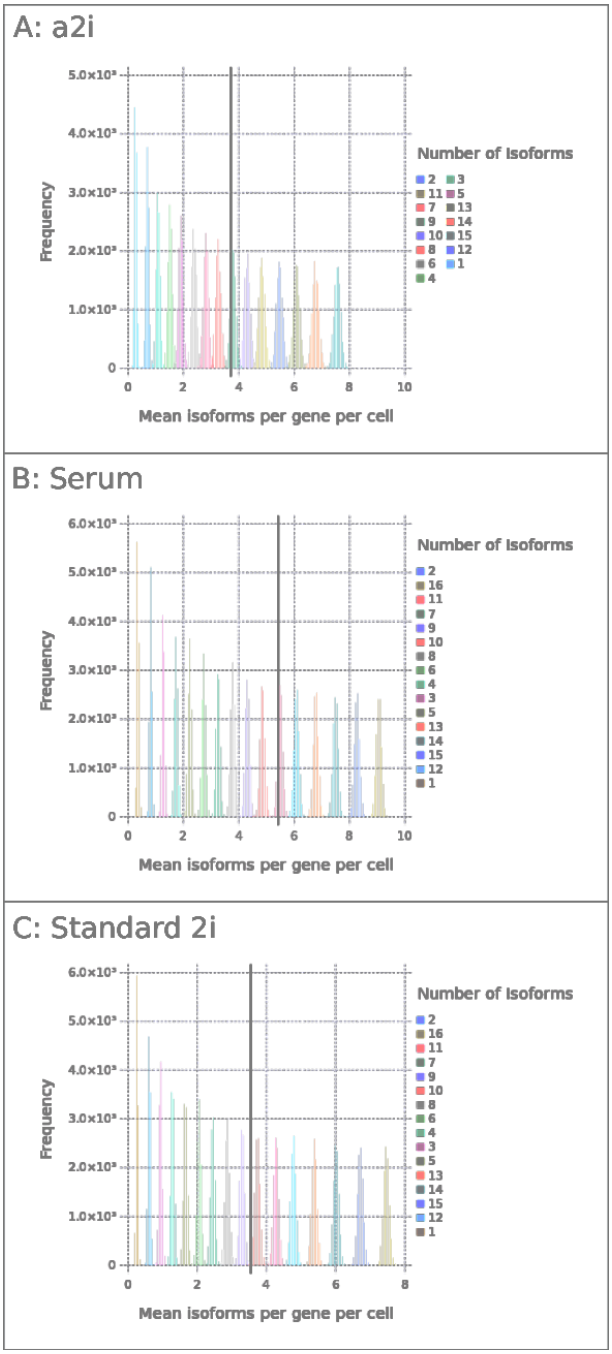


Figure 3.8: Simulation results for Jarid2 gene in mESCs cultured in **A** a2i culture conditions, **B** serum and **C** standard 2i culture conditions. The vertical black line on each plot represents the mean number of isoforms detected per cell in the real data.

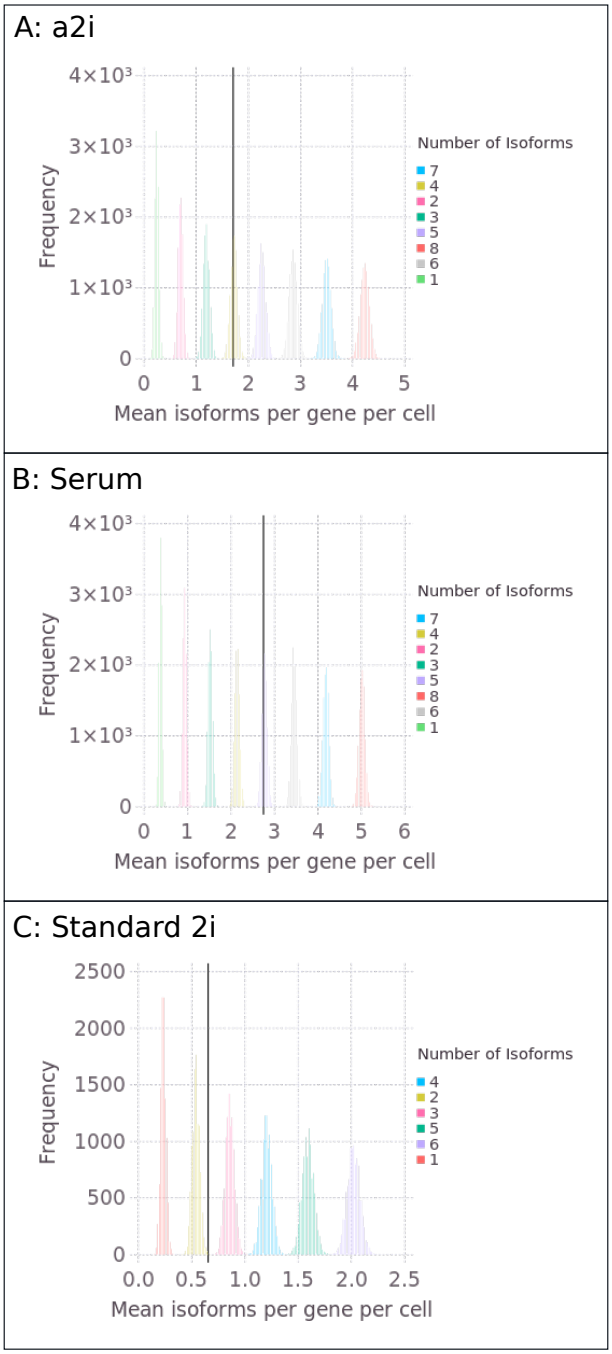


Figure 3.9: Simulation results for Myc gene in mESCs cultured in **A** a2i culture conditions, **B** serum and **C** standard 2i culture conditions. The vertical black line on each plot represents the mean number of isoforms detected per cell in the real data.

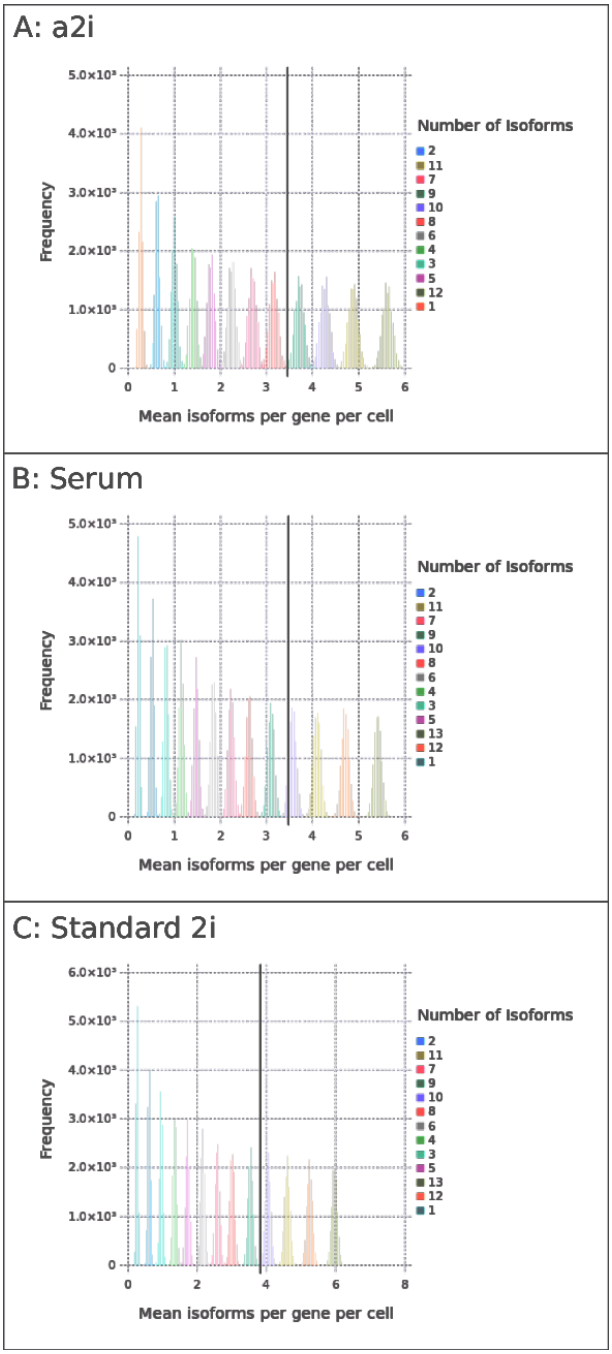


Figure 3.10: Simulation results for Stat3 gene in mESCs cultured in **A** a2i culture conditions, **B** serum and **C** standard 2i culture conditions. The vertical black line on each plot represents the mean number of isoforms detected per cell in the real data.

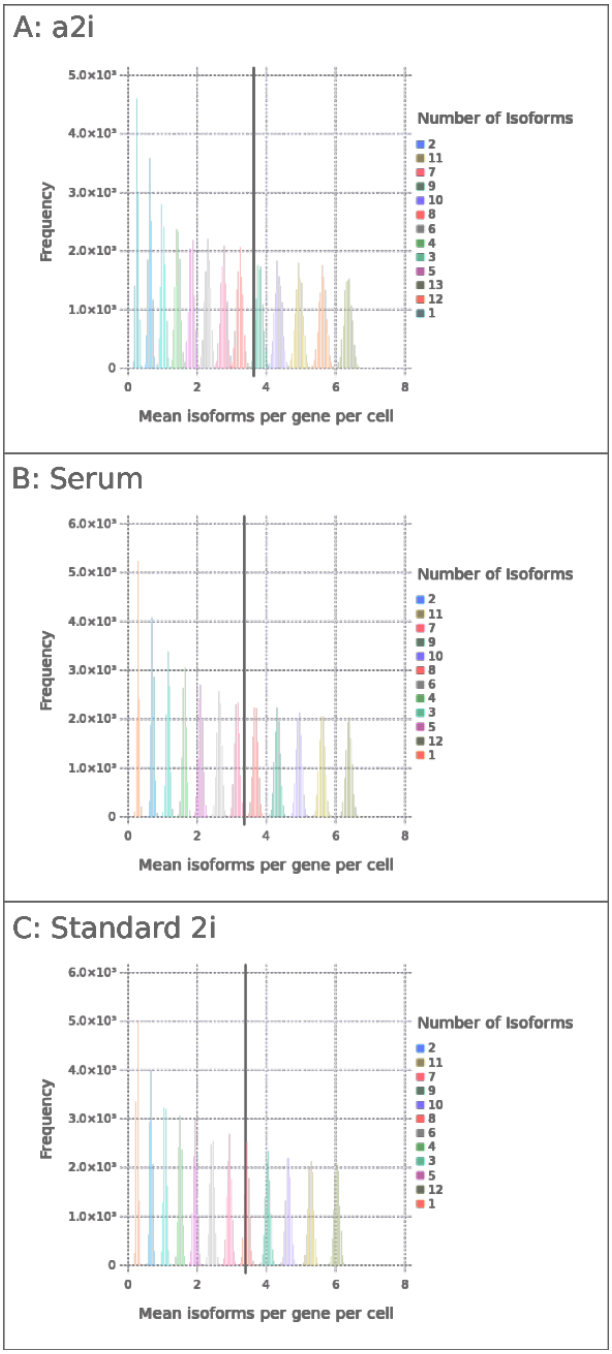


Figure 3.11: Simulation results for Tcf3 gene in mESCs cultured in **A** a2i culture conditions, **B** serum and **C** standard 2i culture conditions. The vertical black line on each plot represents the mean number of isoforms detected per cell in the real data.

### 3.2.3 My novel simulation approach makes unlikely predictions.

An ideal means of validating my model's predictions would be to carry out smFISH to resolve the number of isoforms produced in individual cells. However, resolving between isoforms using smFISH is not trivial, and the experiments would take some time to carry out. Before arranging such experiments, I decided to carry out additional bioinformatics experiments to search for further evidence of differential splicing between culture conditions.

If differential splicing is occurring between culture conditions, I would expect the proportion of counts allocated to each isoform to differ between conditions. To test whether this is the case for our three candidate genes, I have plotted the number of counts allocated to each isoform from Tbx3 (Figure 3.12 & 3.13), Klf4 (Figure 3.14 & 3.15) and Pou5f1 (Figure 3.16 & 3.17). Although the number of counts allocated to each isoform does appear to systematically differ between culture condition, reflecting differential gene expression and/or library size differences, I can see no evidence that the number of isoforms produced differs between culture conditions. For example, in Figure 3.12, more counts are systematically mapped to Tbx3 in the a2i and standard 2i culture conditions compared to serum. However, there is good evidence of expression of three isoforms (ENSMUST00000202034.1, ENSMUST00000121021.7 and ENSMUST0000018748.8) in all three culture conditions and little detection of ENSMUST00000079719.10, ENSMUST00000145647.1 and ENSMUST00000154680.1. This appears to contradict my model's prediction that 3-4 Tbx3 isoforms are expressed in a2i, 1-2 isoforms are expressed in serum and 4-5 isoforms are present in standard 2i. It is of course possible that more isoforms are produced across a group of cells than in individual cells, but my model's predictions that more isoforms are produced in individual cells than can be detected across a group of cells are less plausible. In Figure 3.13, I investigated how many Tbx3 isoforms could be detected in the matched bulk RNA-seq data, and find good evidence that three isoforms are produced in all three culture conditions, further calling into question my model's prediction that more than three isoforms are produced

in some cells. Similarly, in Figures 3.14 - 3.17 there is good evidence of expression of one isoform in all three culture conditions for Klf4 and Pou5f1 in both the bulk and scRNA-seq data, despite my model predicting substantially more isoforms being produced in individual cells. Based on these results, I am forced to conclude that as my model currently stands, it cannot make reliable predictions about the number of isoforms produced in individual cells.

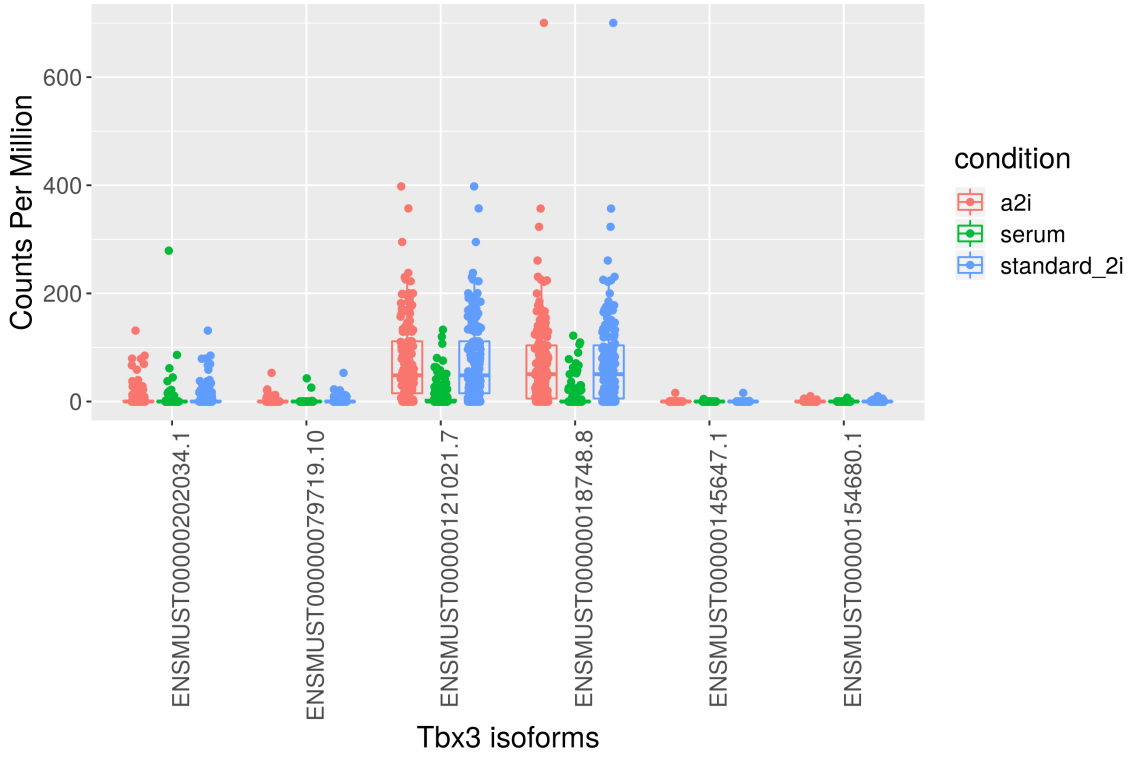


Figure 3.12: Boxplots of scRNA-seq counts mapping to each Tbx3 isoform in the three culture conditions. Points and boxplots are coloured by culture condition.

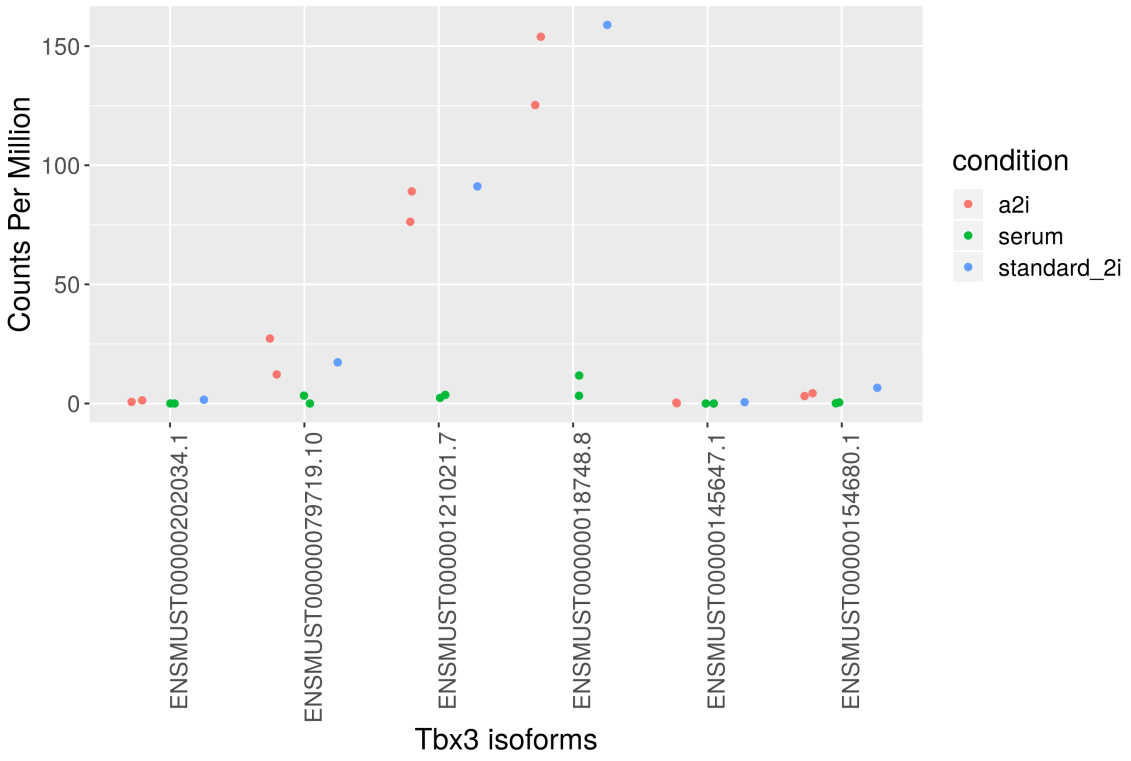


Figure 3.13: Plots of matched bulk RNA-seq counts mapping to each Tbx3 isoform in the three culture conditions. Points are coloured by culture condition.

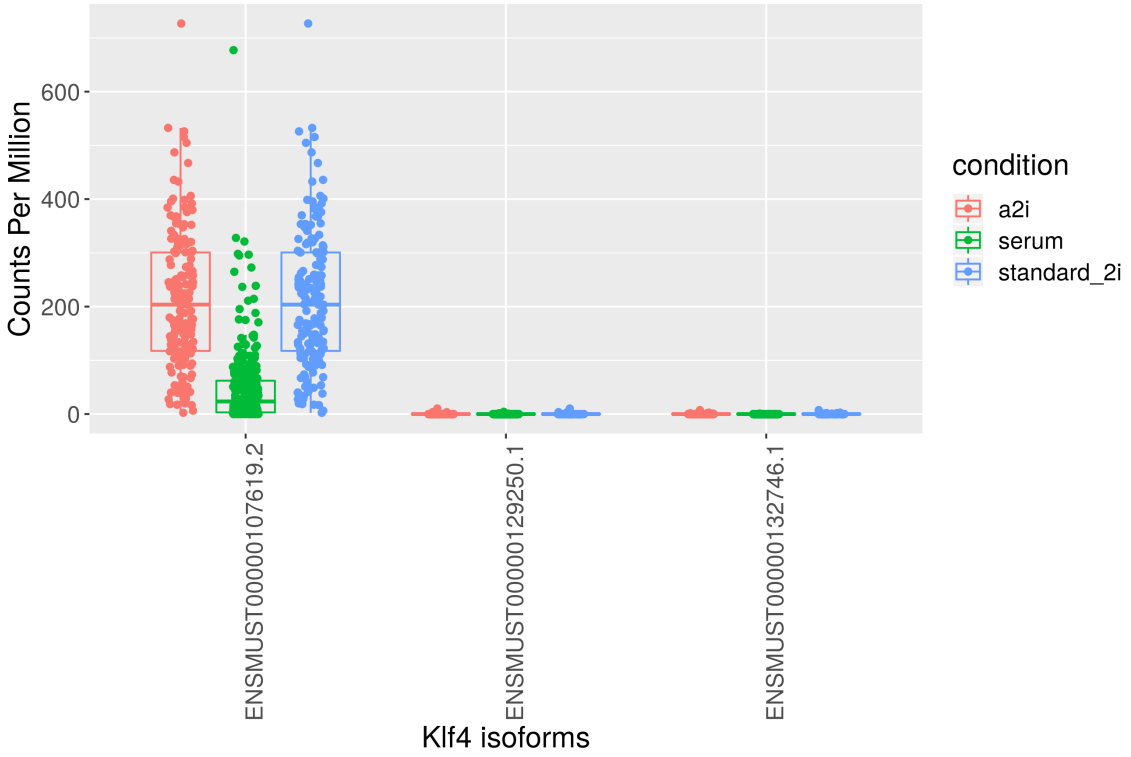


Figure 3.14: Boxplots of scRNA-seq counts mapping to each Klf4 isoform in the three culture conditions. Points and boxplots are coloured by culture condition.

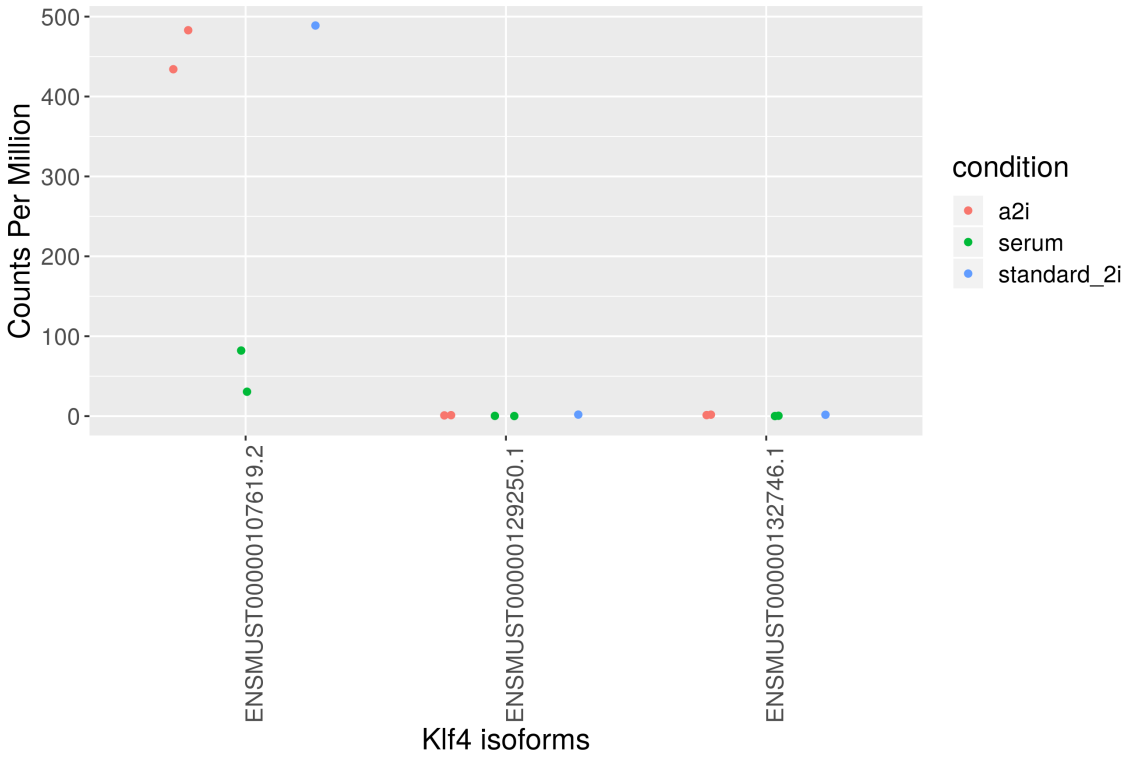


Figure 3.15: Plots of matched bulk RNA-seq counts mapping to each Klf4 isoform in the three culture conditions. Points are coloured by culture condition.

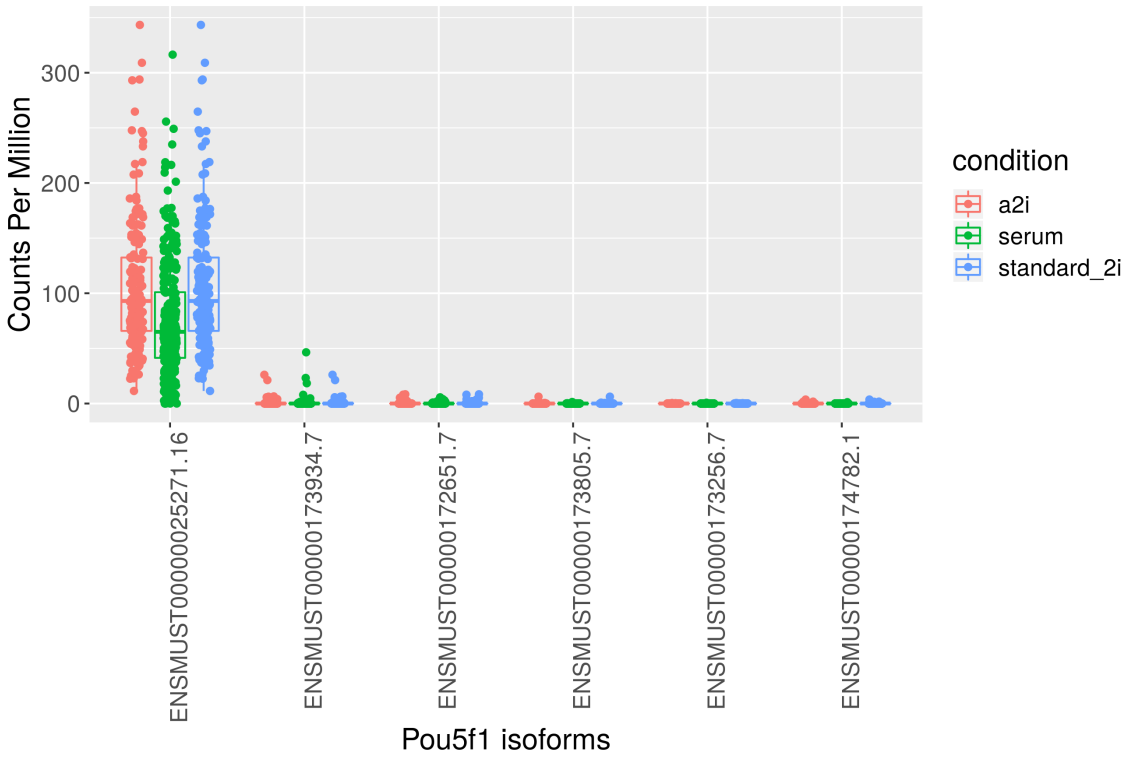


Figure 3.16: Boxplots of scRNA-seq counts mapping to each Pou5f1 isoform in the three culture conditions. Points and boxplots are coloured by culture condition.

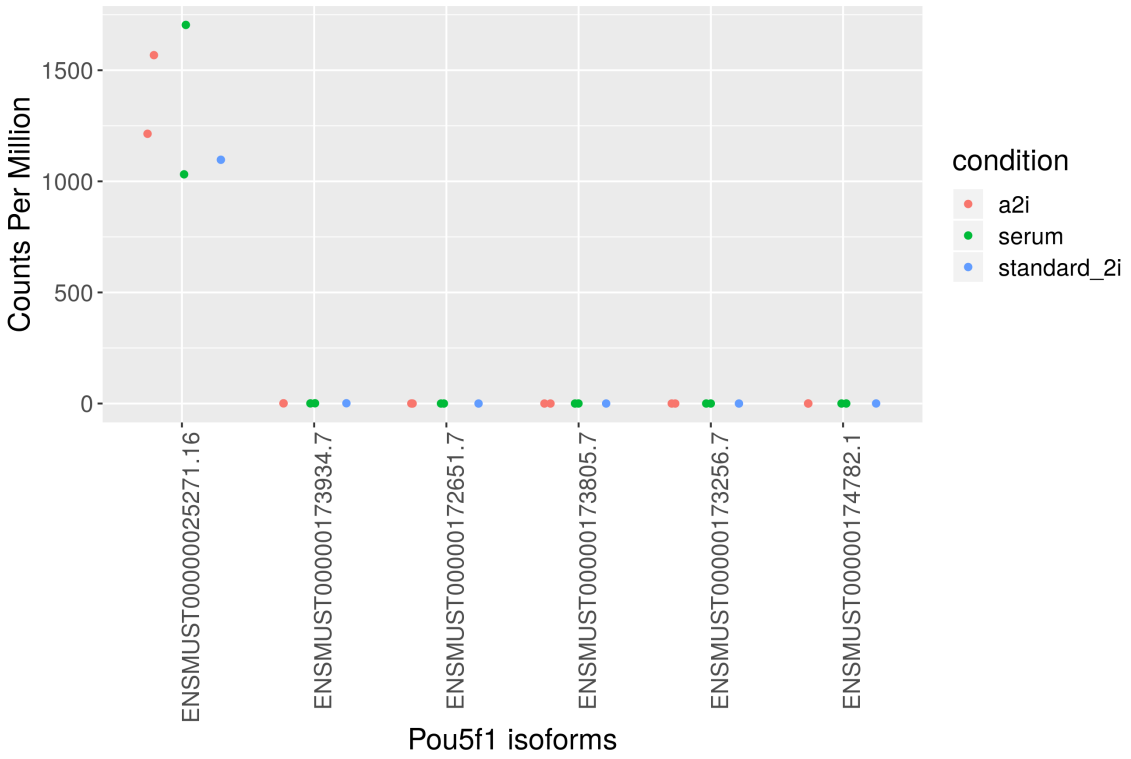


Figure 3.17: Plots of matched bulk RNA-seq counts mapping to each Pou5f1 isoform in the three culture conditions. Points are coloured by culture condition.

## 3.3 Discussion

I began this chapter by investigating how many isoforms are detected in scRNA-seq for genes where two isoforms are detected in matched bulk RNA-seq in a dataset of mESCs and quiescent B lymphocytes. I found that it is rare to detect both isoforms and common to fail to detect gene expression at all in many cells. However, both isoforms were detected for some genes, and it was more common to detect both isoforms in individual cells if the parent gene was highly expressed. Without a clear idea of how best to correct for dropouts, it is impossible to state to what extent these observations reflect biological facts, and to what extent they are consistent with every cell producing both isoforms but few isoforms being detected due to dropouts.

I attempted to correct for this by developing a simulation based model for predicting how many isoforms are produced from a gene of interest in individual cells. I explicitly simulated dropouts using a popular model for dropout probability (Andrews and Hemberg, 2018a). However, my model made some questionable predictions, most notably that more isoforms were produced per cell than could be detected across all cells. This is not entirely impossible if the rate of dropouts is so high that we almost entirely fail to detect some expressed isoforms across a population of cells. However, that we detect fewer isoforms than my model predicts are produced in the matched bulk RNA-seq data makes the dropout hypothesis unlikely. I believe the most likely explanation is that my model is not making accurate predictions, suggesting that the model needs to be refined. Identifying how best to refine my model is not trivial. Possible issues with my model include the following:

1. My model might be sensitive to library size.
2. I might not be modelling dropouts sufficiently accurately.
3. In reality, different cells might produce different numbers of isoforms. This behaviour is not captured in my simulations.
4. I might not be modelling quantification errors sufficiently accurately.

5. I might not be modelling the process of isoform choice within individual cells accurately.

Some of these issues are more straightforward to address than others. Incrementally tweaking my model, and probably making it increasingly complex, is likely to eventually generate a model that generates plausible seeming predictions. However, whether the model that is eventually generated has any actual predictive power or biological relevance is uncertain. More generally, I am concerned that tweaking my model until it produces results that I like is not a very scientific way to develop a predictive model for the number of isoforms produced per gene per cell. A small number of studies have used smFISH to investigate the number of isoforms produced per gene per cell (Velten et al., 2015; Ciolli Mattioli et al., 2019; Waks et al., 2011), but this is not a large or comprehensive enough ground truth dataset to facilitate a machine learning approach to solve this problem. If more smFISH data resolving the number of isoforms produced per gene per cell were generated, a machine learning approach might become more feasible.

Unfortunately, I lack the skills and time necessary to create a large smFISH dataset, so a different approach is required. The major issue underlying my list of possible modelling issues is that as a field, we do not have a sophisticated understanding of the technical noise that is present in scRNA-seq data. We are aware that technical noise, in the form of dropouts, quantification errors, batch effects, PCR amplification bias and other sources, exists. However, the extent to which these different sources of technical noise confound alternative splicing analyses and how best to correct for these confounders is not known. In the next chapter, I attempt to start solving this problem. I take another simulation based approach in which I vary the amount of technical noise in scRNA-seq data and investigate what the biggest confounders are when studying alternative splicing. After identifying the major confounders, I propose solutions that could allow the field to overcome these confounders and enable accurate alternative splicing analyses using scRNA-seq.

## 3.4 Conclusions

Intriguing patterns of isoform expression exist in scRNA-seq data, but establishing to what extent these patterns are biologically real requires a more sophisticated understanding of the technical noise that exists in scRNA-seq. I will investigate the extent to which technical noise confounds splicing analyses in scRNA-seq in the next chapter.

# 4

# Obstacles to Detecting Isoforms Using Full-Length scRNA-seq Data.

*Negative results are just what I want. They're just as valuable to me as positive results. I can never find the thing that does the job best until I find the ones that don't.*

– Attributed to Thomas Edison

## Introduction

Thus far in my thesis, I have established that although software to accurately quantify isoforms using scRNA-seq data exists, it remains challenging to analyse splicing using scRNA-seq data due to uncertainty over how to correct for the large amounts of technical noise present. To begin to understand how to correct for confounding factors in scRNA-seq to enable splicing analyses, we first need to establish what factors are actually confounding our analyses. For example, I hypothesised in the previous chapter that technical dropouts could be confounding my splicing experiments. Whilst it seems reasonable to suppose that scRNA-seq's low capture efficiency might

prevent us from accurately detecting the number of isoforms expressed in individual cells, to the best of my knowledge this hypothesis has never been systematically tested. Consequently, the extent to which dropouts confound splicing analyses using scRNA-seq data is not known. The same is true for other sources of technical noise associated with scRNA-seq.

In this chapter, I take a novel approach to investigate what confounders are present when studying splicing using scRNA-seq. I take real scRNA-seq datasets and select genes for which four isofoms are detected. I then use these genes to simulate the following four scenarios: 1) all cells express one isoform per gene per cell, 2) all cells express two isoforms per gene per cell, 3) all cells express three isoforms per gene per cell, and 4) all cells express four isoforms per gene per cell. Importantly, in each scenario I explicitly simulate dropout events and quantification errors. I then use the simulated output of each scenario to ask two questions. Firstly, to what extent is it possible to distinguish between these global differences in alternative splicing using scRNA-seq? And secondly, what should be done to enable more accurate splicing analysis with scRNA-seq? I find that dropouts are a major confounder when attempting to study alternative splicing using scRNA-seq, whereas quantification errors are a much lesser confounder. I find that different models of isoform choice meaningfully impact on my simulation results, indicating that isoform choice may need to be considered in future splicing analyses.

The work presented in this chapter has been released on bioRxiv, consequently some passages have been quoted verbatim from the following source: (Westoby et al., 2019). Additionally, some figures have been reproduced from the aforementioned source. At the time of writing, this work is in revision for Genome Biology.

## 4.1 Results

A detailed description of my simulation approach can be found in the Methods chapter, a brief description is given here for convenience. My approach for the first scenario, in which I simulate one isoform being expressed per gene per cell, is to first identify genes for which the expression of exactly four isoforms is detected in a real

scRNA-seq dataset. The reasoning for selecting genes which express four isoforms is that four isoforms is a sufficiently large number of isoforms to be sufficient to study a range of splicing behaviours. At the same time, four isoforms is sufficiently few isoforms that there are not substantial concerns about an increase in the quantification error rate due to the presence of many isoforms sharing a high degree of sequence identity.

In the second step, I randomly select one isoform based on a plausible model of isoform choice for the first of the genes in the first cell in the simulated dataset. For my default model of isoform choice, I choose the isoform based on a model of alternative splicing described by Hu et al. (Hu et al., 2017). Third, I simulate dropouts based on a Michaelis-Menten model described by Andrews and Hemberg (Andrews and Hemberg, 2018a). Fourth, I simulate quantification errors based on isoform detection error estimates based on work by Westoby et al. (Westoby et al., 2018b). I repeat these four steps for every four isoform gene and cell in our simulated dataset, then calculate the mean number of isoforms detected for that gene per cell. The entire process described above is one complete simulation. I run 100 simulations for each of our four scenarios, where each scenario corresponds to one, two, three or four isoforms being expressed per gene per cell. I then plot the distributions of the mean number of isoforms detected per gene per cell for each scenario. A schematic of my simulation approach is displayed in Figure 4.1. Negative control models, in which my simulations are repeated but with no dropouts and/or quantification errors are simulated, can be found in Figures 4.2 - 4.4.

In Figures 4.5 & 4.6, I apply my simulation approach to a dataset of H1 and H9 human embryonic stem cells (hESCs) (Bacher et al., 2017). In this dataset, each cell's cDNA was split into two groups and sequenced at two different sequencing depths, enabling me to directly compare our simulation results at different sequencing depths without biological confounders. One group was sequenced at approximately 1 million reads per cell and the other group at approximately 4 million reads per cell on average. My simulation results for the two H1 groups are compared side by side in Figure 4.5A. My simulation results for the two H9 groups are shown in Figure 4.6A. scRNA-seq experiments have been found to saturate in terms of the number of genes

detected per cell at approximately 1 million reads per cell (Svensson et al., 2017; Ziegenhain et al., 2017). However, I observe differences in the number of isoforms detected per gene per cell at 1 and 4 million reads per cell, indicating that the saturation depth may differ for gene and isoform level analyses. Next, I calculate the fraction of overlap between the isoforms expressed in the ground truth and the isoforms detected as expressed in our simulations. In Figures 4.5B & 4.6B, I show the distributions of the mean fraction of overlap for each gene. I will refer to the each gene's mean fraction of overlap between isoforms expressed in the ground truth and isoforms detected as expressed as the 'overlap fraction' hereafter in the text. The mean overlap fraction is consistently higher at 4 million reads per cell compared to at 1 million reads per cell, indicating that our ability to accurately detect isoforms is improved at higher sequencing depths.

Figures 4.5 & 4.6 illustrate some of the difficulties associated with splicing analysis in scRNA-seq. At both sequencing depths, the distributions of the observed mean number of isoforms per gene per cell are shifted to the left of their true value. In addition, the highest mean overlap fraction observed is less than 0.8, indicating that even in a best case scenario, we fail to detect over 20% of the isoforms expressed in the ground truth. These effects are less extreme, but still present, for the group sequenced at approximately 4 million reads per cell compared to the group sequenced at 1 million reads per cell. This is consistent with the hypothesis that sequencing at higher depth reduces the extent to which isoform number is underestimated. However, even at approximately 4 million reads per cell our simulations suggest that scRNA-seq substantially underestimates the mean number of isoforms per gene per cell for almost all genes. A naive analysis of these two datasets would most likely underestimate the number of isoforms expressed per gene per cell. This casts doubt on the biological relevance of previous observations suggesting only one isoform was typically produced per gene per cell, although admittedly the sequencing depth per cell was generally much greater than 4 million reads per cell in those studies (for example, Shalek et al. sequenced approximately 27 million reads per cell (Shalek et al., 2013)).

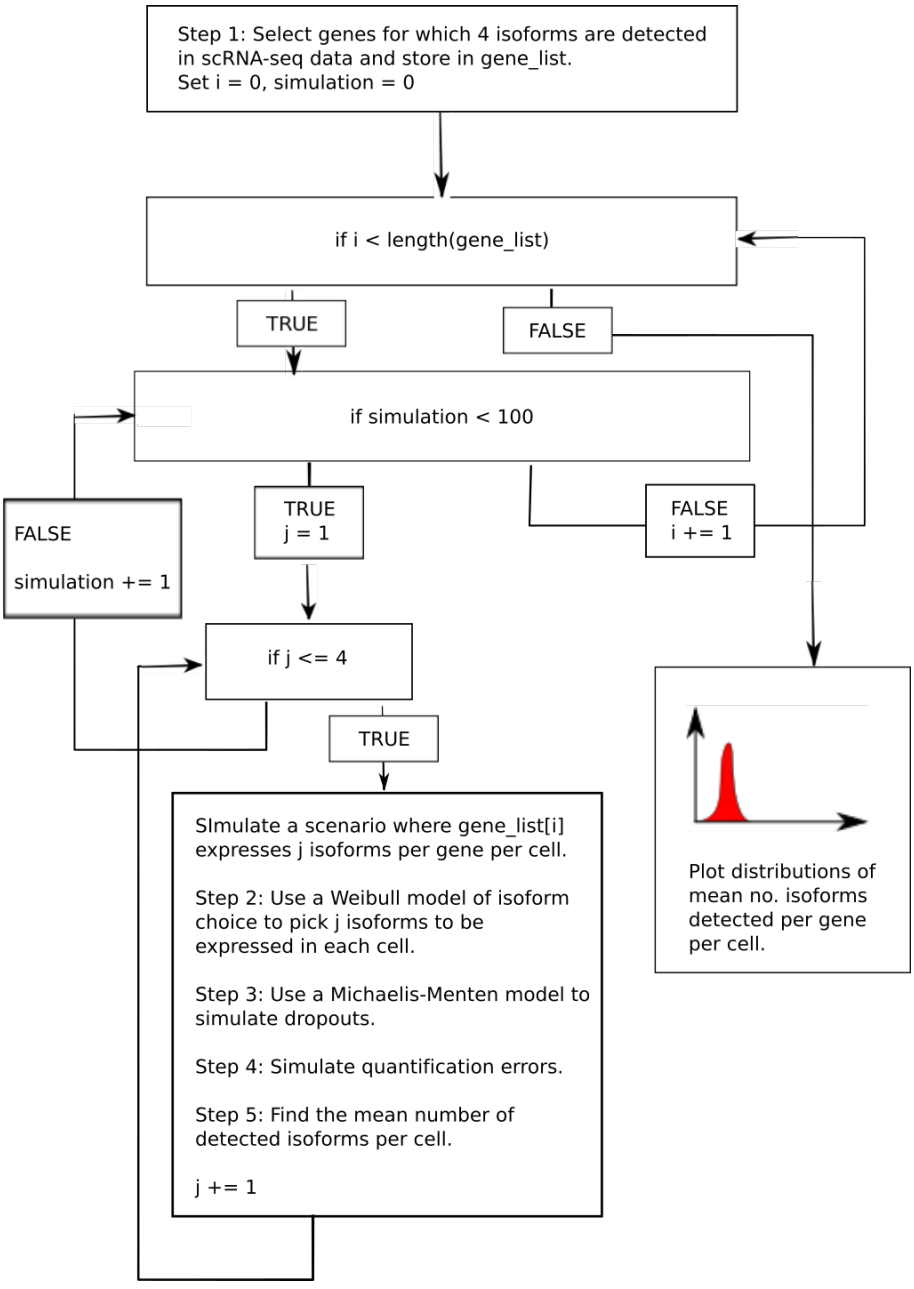


Figure 4.1: Schematic of our simulation approach.

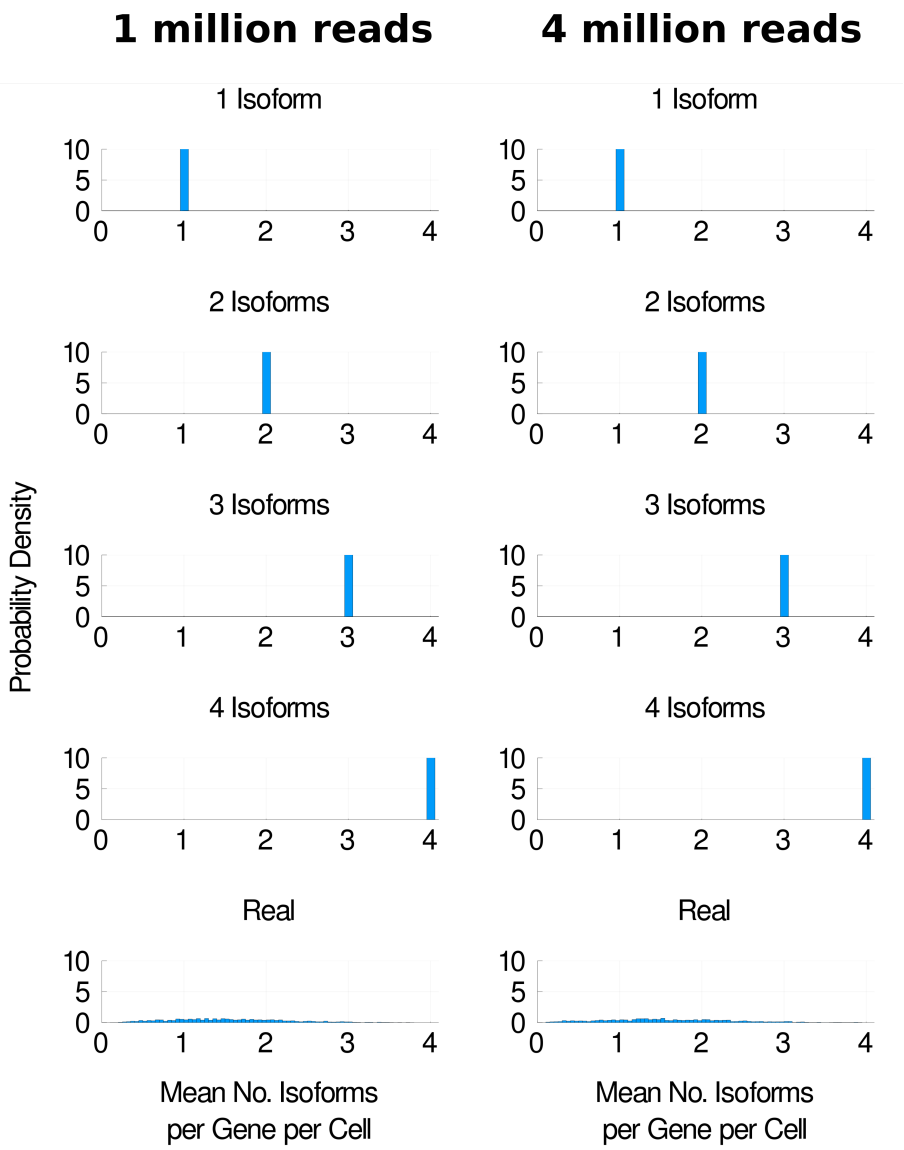


Figure 4.2: Negative control model for H1 hESCs. In the simulation results displayed, no dropouts or quantification errors were simulated. The simulation procedure was otherwise unchanged.

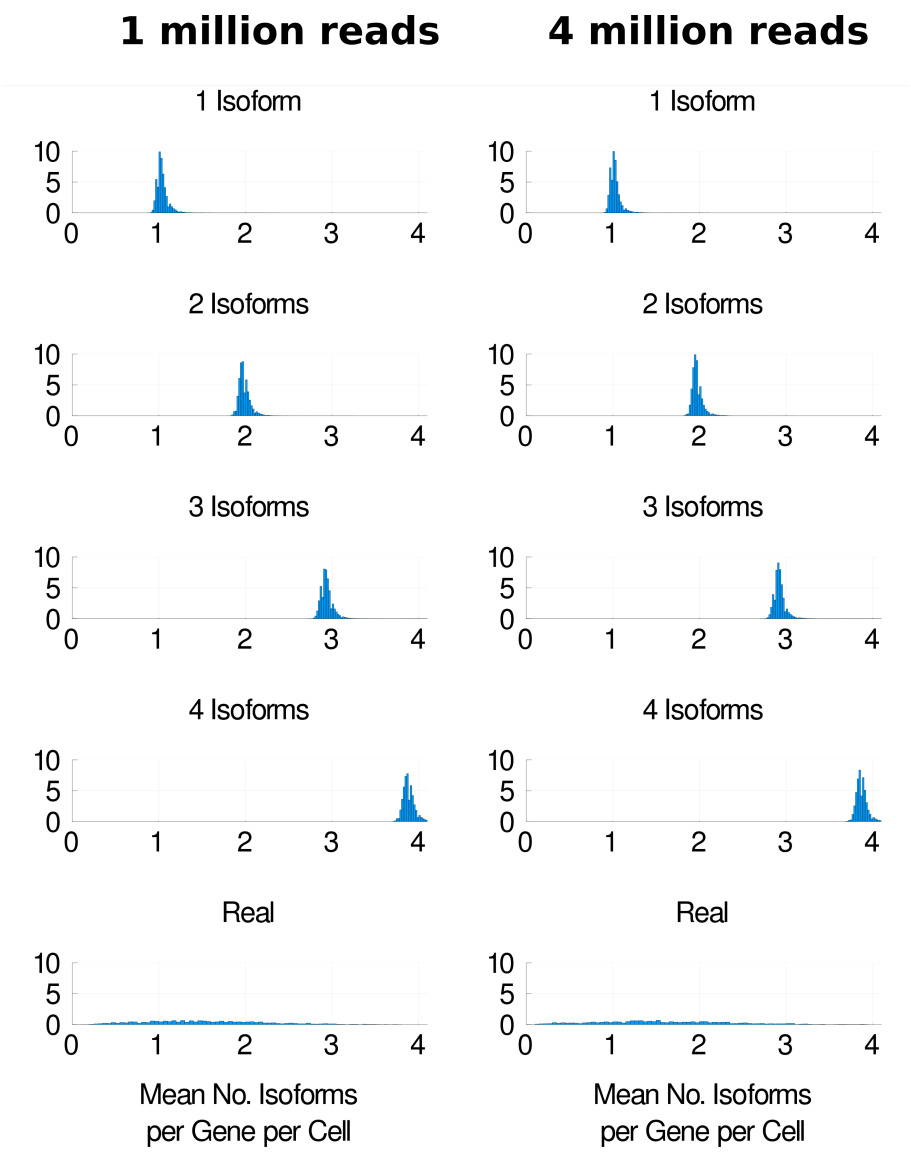


Figure 4.3: Negative control model for H1 hESCs. In the simulation results displayed, no dropouts were simulated. The simulation procedure was otherwise unchanged.

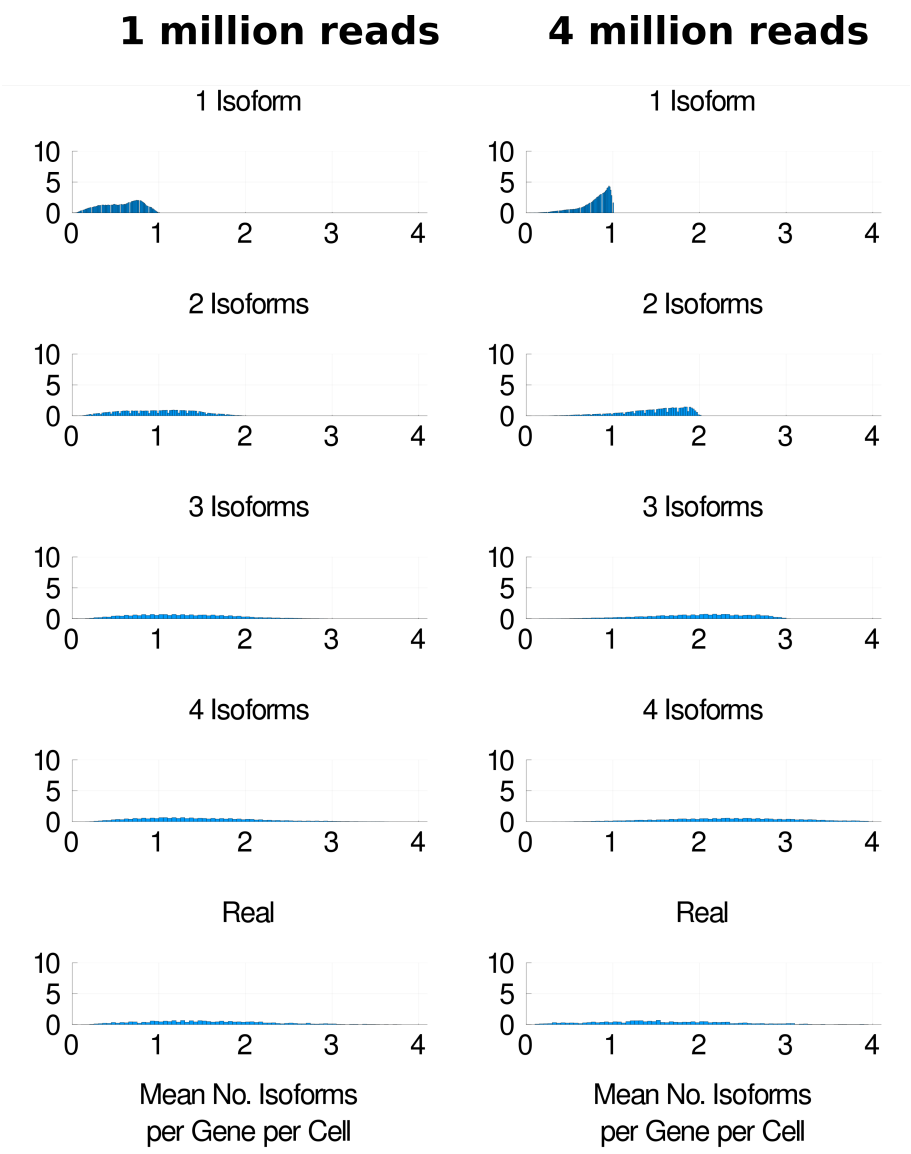


Figure 4.4: Negative control model for H1 hESCs. In the simulation results displayed, no quantification errors were simulated. The simulation procedure was otherwise unchanged.

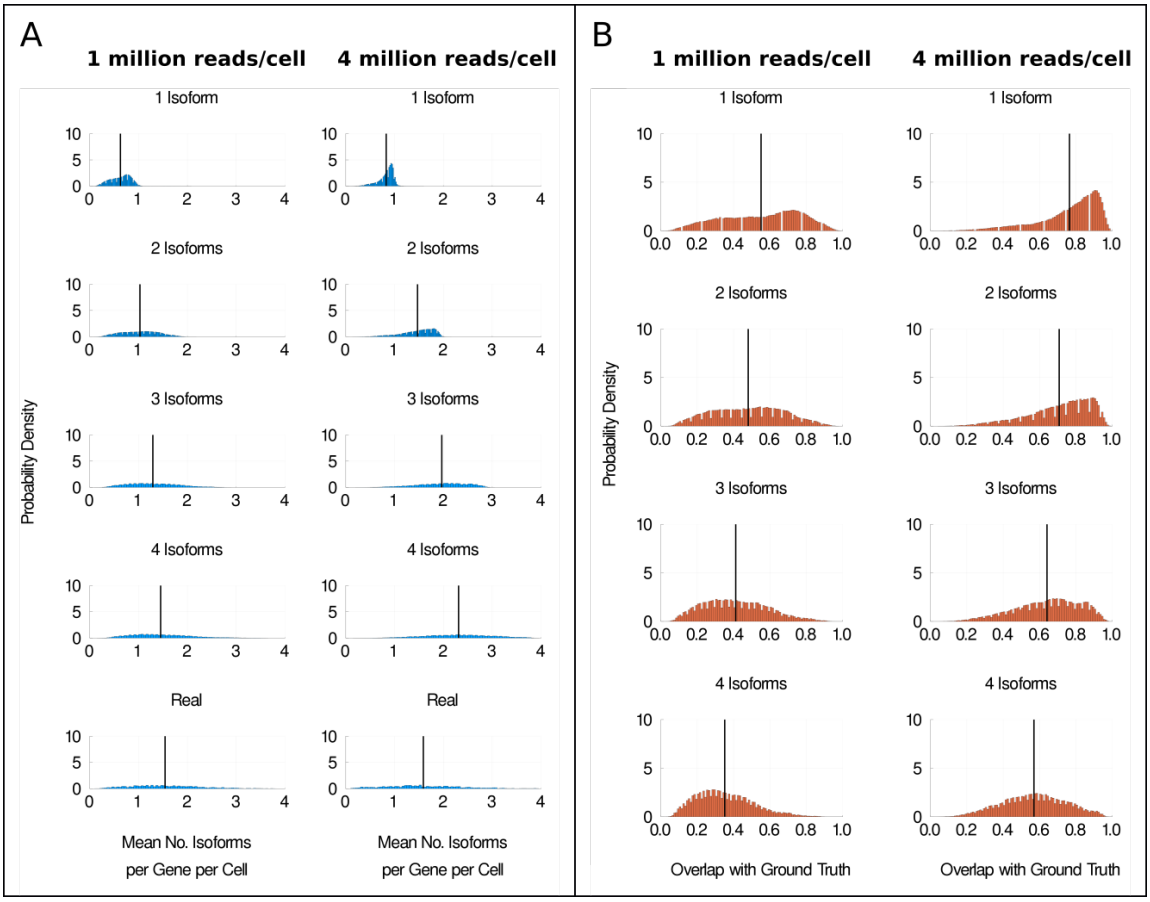


Figure 4.5: The effect of sequencing depth on isoform detection. **A** Distributions of the mean number of isoforms detected per gene per cell for H1 hESCs whose cDNA was split and sequenced at approximately 1 million reads per cell or 4 million reads per cell on average. **B** Distributions of the overlap fraction. Black vertical lines represent the mean value of the distributions.

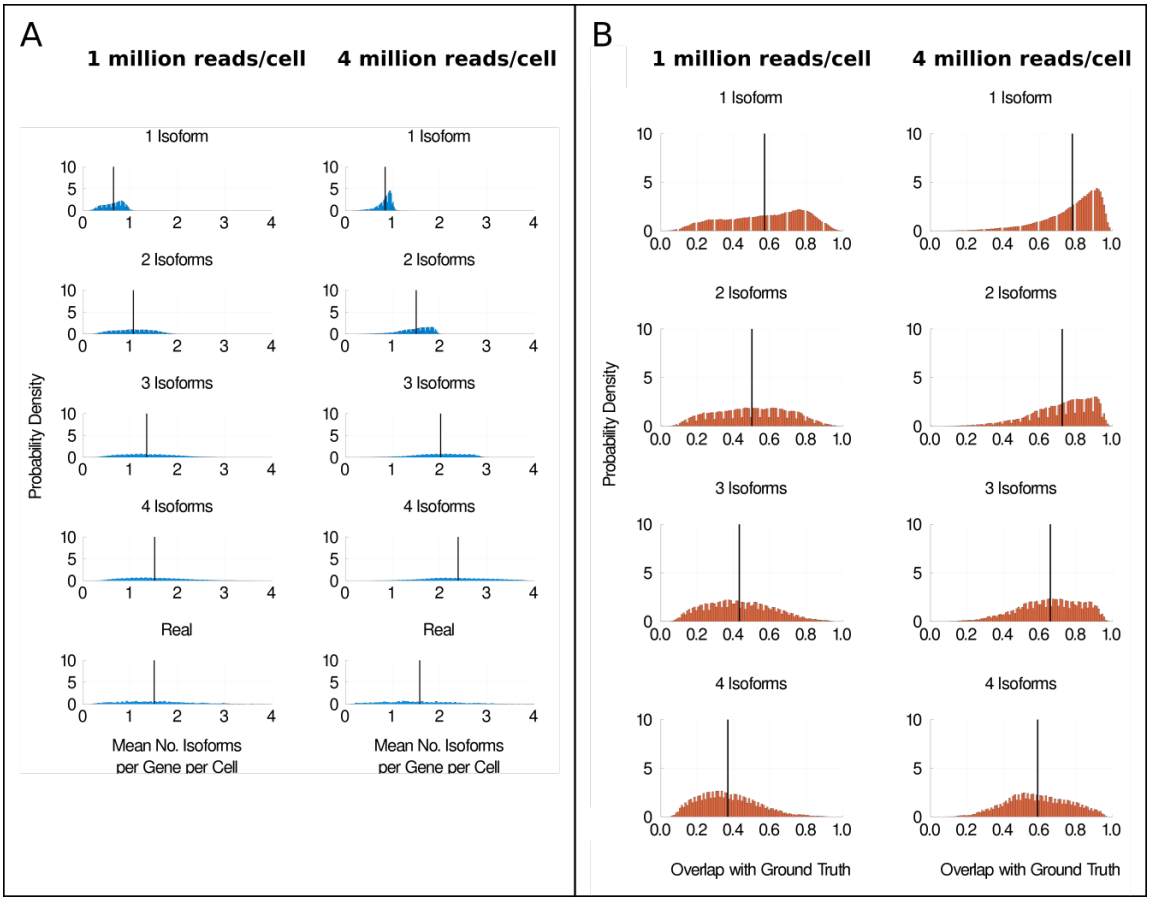


Figure 4.6: The effect of sequencing depth on isoform detection. **A** Distributions of the mean number of isoforms detected per gene per cell for H9 hESCs whose cDNA was split and sequenced at approximately 1 million reads per cell or 4 million reads per cell on average. **B** Distributions of the overlap fraction. Black vertical lines represent the mean value of the distributions.

One hypothesis for why our ability to detect isoforms increases with increased sequencing depth is that the rate of dropouts is reduced. In Figures 4.7A & 4.8A, I investigate this hypothesis by plotting the distribution of the probabilities of dropout for each isoform (p(dropout)), as estimated using the Michaelis-Menten equation (Andrews and Hemberg, 2018a) (see Methods chapter). I find that the distribution is skewed towards high probabilities of dropout for the group sequenced at around 1 million reads per cell. In contrast, the distribution for the group sequenced at around 4 million reads per cell is more skewed towards low probabilities of dropouts. This demonstrates that my estimated dropout probabilities are different at the two sequencing depths, as expected.

Overall, the data in Figures 4.5 & 4.7A and in Figures 4.5 & 4.8A support the hypothesis that when the rate of technical dropouts decreases, the accuracy of isoform number estimation increases. However, as the dataset was only sequenced at two depths, I only have two data points available to investigate my hypothesis. To extend my investigation, I assume that the distributions of dropout probabilities observed in Figures 4.7A & 4.8A can be modelled as Beta distributions. The Beta distribution is parameterised by two values, $\alpha$ and $\beta$, and I find that it approximates the probability distributions well (see bottom panels of Figures 4.7A & 4.8A). Therefore, I select five values of $\alpha$ and $\beta$ that generate differently shaped dropout distributions, as shown in Figures 4.7B & 4.8B. I then perform five further simulation experiments. In each simulation experiment, I sample our dropout probabilities from one of our Beta distributions. The results of these experiments are shown in Figures 4.7C & D and in Figures 4.8C & D.

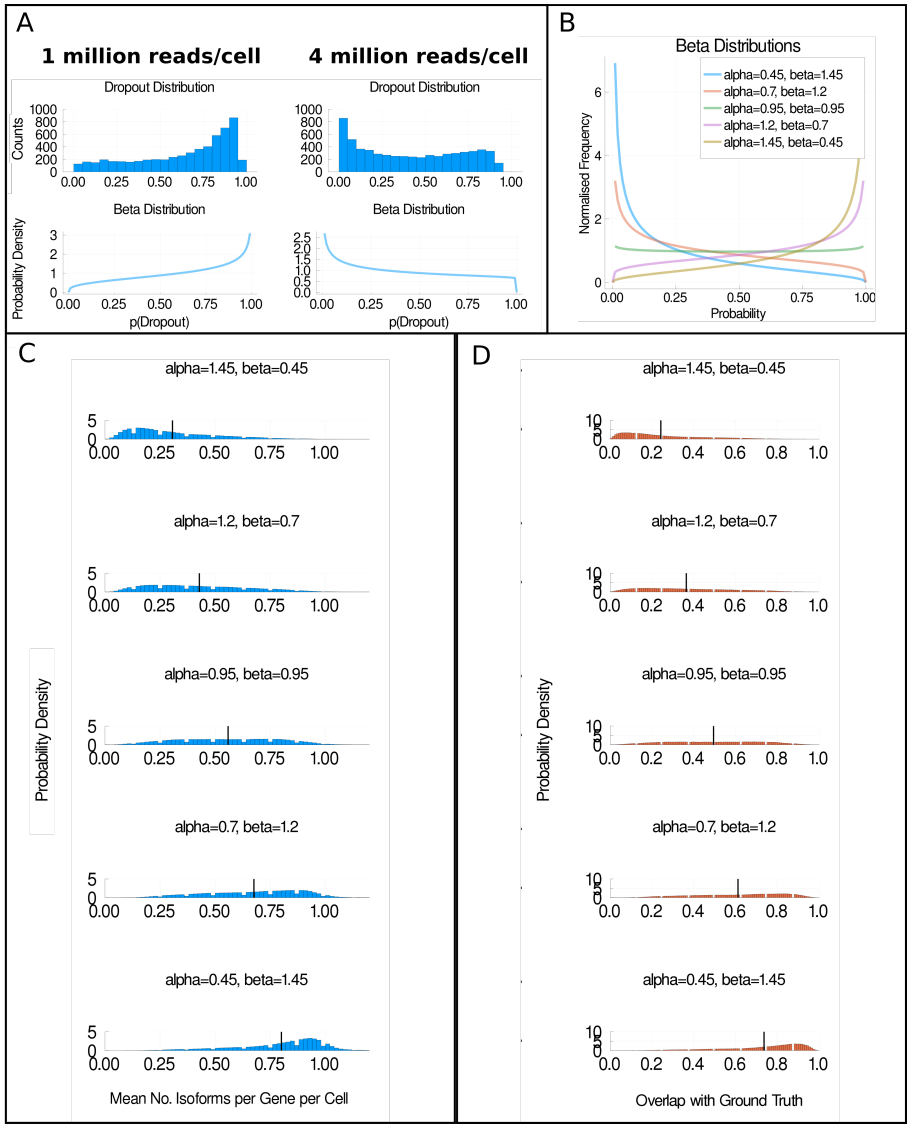


Figure 4.7: The impact of dropouts on isoform detection. **A** shows the distribution of the probabilities of dropouts (p(Dropout)) in each group of H1 hESCs and an approximation of these distributions using a Beta distribution. At 1 million reads per cell, $\alpha = 1.31$ and $\beta = 0.74$ in the approximated Beta distribution. At 4 million reads per cell, $\alpha = 0.72$ and $\beta = 1.03$ in the approximated Beta distribution. **B** shows five Beta Distributions from which dropout probabilities were sampled from in the simulations used to generate **C** and **D**. In **C**, the distribution of the mean number of isoforms detected per gene per cell is shown for simulations in which one isoform was produced per gene per cell. Each plot corresponds to a simulation in which dropout probabilities were sampled from one of the distributions shown in **B**. **D** shows the overlap fraction for each simulation. Plots shown in **C** & **D** are for H1 hESCs sequenced at 4 million reads per cell. Black vertical lines represent the mean value of the distributions.

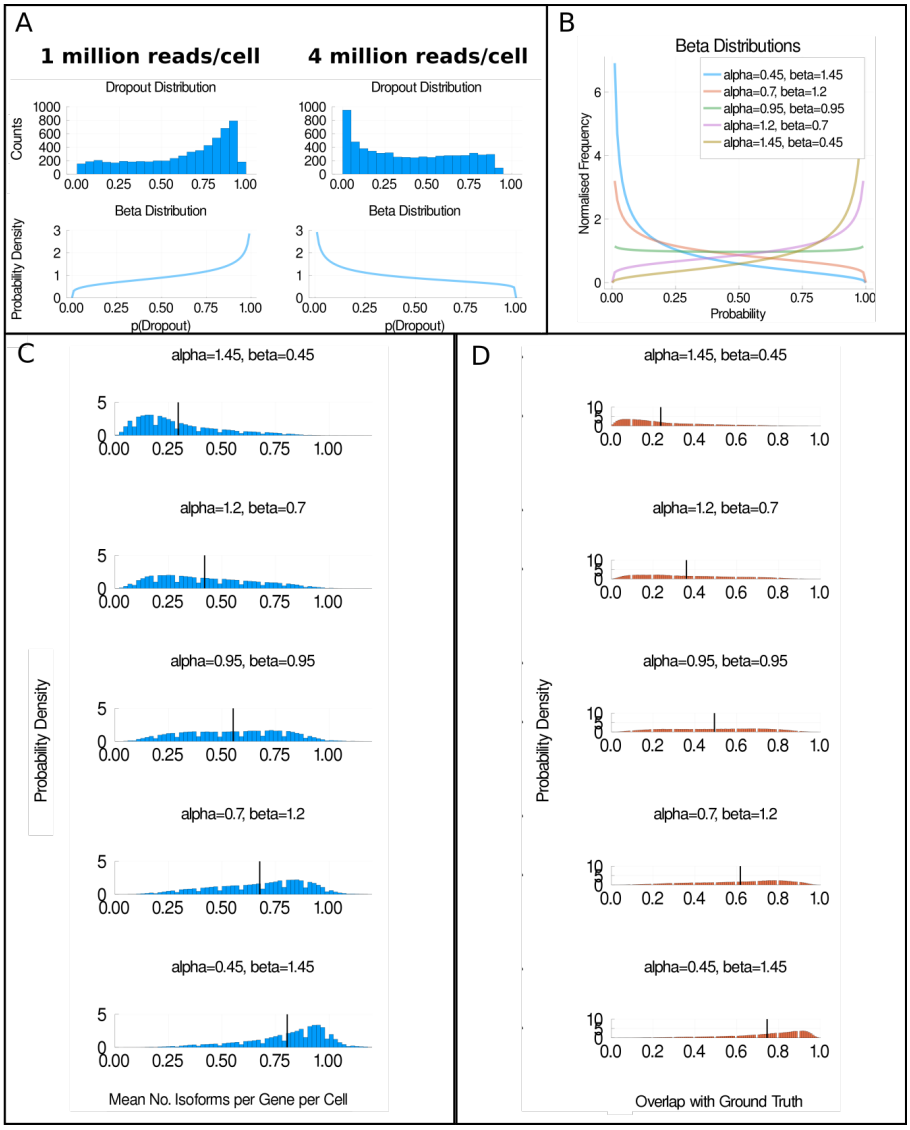


Figure 4.8: The impact of dropouts on isoform detection. **A** shows the distribution of the probabilities of dropouts (p(Dropout)) in each group of H9 hESCs and an approximation of these distributions using a Beta distribution. At 1 million reads per cell, $\alpha = 1.31$ and $\beta = 0.74$ in the approximated Beta distribution. At 4 million reads per cell, $\alpha = 0.72$ and $\beta = 1.03$ in the approximated Beta distribution. **B** shows five Beta Distributions from which dropout probabilities were sampled from in the simulations used to generate **C** and **D**. In **C**, the distribution of the mean number of isoforms detected per gene per cell is shown for simulations in which one isoform was produced per gene per cell. Each plot corresponds to a simulation in which dropout probabilities were sampled from one of the distributions shown in **B**. **D** shows the overlap fraction with the ground truth for each simulation. Plots shown in **C** & **D** are for H9 hESCs sequenced at 4 million reads per cell.

In Figures 4.7C & 4.8C, I show the mean detected number of isoforms per gene per cell for the scenario where each gene produces one isoform per gene per cell. As I move from the top to the bottom of Figures 4.7C & 4.8C, the value of $\alpha$ decreases, corresponding to scenarios where the probability of dropout is more frequently close to zero. As $\alpha$ decreases, the distributions of mean detected isoforms per gene per cell shift further to the right and closer to the true number of isoforms produced per cell. In Figures 4.7D & 4.8D, I find that the mean overlap fraction increases as $\alpha$ decreases, corresponding to the mean probability of dropout decreasing. I conclude from Figures 4.7 C & D and from Figures 4.8 C & D that reducing the dropout rate would likely improve the accuracy of splicing analyses performed using scRNA-seq.

### 4.1.1 Quantification errors are a relatively minor obstacle to studying alternative splicing

A benchmark of isoform quantification softwares in full length coverage mouse scRNA-seq datasets found that the error rate of many software tools was low and comparable to bulk RNA-seq (Westoby et al., 2018b). This is encouraging, however it should be noted that the error rate is likely to be substantially higher for non-model organisms with less well annotated genomes than the mouse genome. As isoform quantification is a key step of many scRNA-seq alternative splicing analysis pipelines, it would be beneficial to understand how quantification errors impact our ability to study alternative splicing, both when the error rate is high and when the error rate is low.

As my interest in this study is the detected number of isoforms per gene per cell, I am only interested in quantification errors which lead to changes in the number of isoforms detected. We simulate two types of quantification errors, false positives and false negatives. In this context, a false positive occurs when an isoform is called as expressed by the quantification software when there are no reads from that isoform. Note that this means that if an isoform is expressed in a cell but no reads are captured from it (i.e. a dropout), but the quantification software calls it as expressed, we would define this as a false postive event. A false negative occurs when an isoform is not called as expressed by the isoform quantification software when reads from that

isoform are present. Based on my previous benchmark (Westoby et al., 2018b), I estimate that the probability of false positive events ($pFP$) is around 1% and that the probability of false negative ($pFN$) events is around 4% (see Methods chapter). In my simulations in Figure 4.9, I vary both of these probabilities in the range of 0% to 50% . Figure 4.9A shows how the mean number of isoforms detected per gene per cell distributions changes as the probability of false positives and false negatives alters when every gene expresses one isoform per cell. Importantly, even when the probability of false positives and false negatives is zero, there are many genes for which the mean number of detected isoforms per gene per cell is not equal to one, the true number of expressed isoforms. This indicates that even if a perfect, 100% accurate isoform quantification tool existed, there would still be substantial barriers to studying alternative splicing using scRNA-seq. We suspect that the reason a 100% accurate isoform quantification tool would underestimate the number of isoforms per gene per cell is that isoform quantification tools usually only quantify the reads that are present. Due to the high number of dropouts in scRNA-seq, many expressed isoforms do not generate reads and thus would be called as unexpressed by a 100% accurate isoform quantification tool, leading to an underestimate of the number of isoforms present.

Unsurprisingly, increasing the probability of false positives causes an increase in the mean number of detected isoforms, whilst increasing the probability of false negatives causes the mean number of detected isoforms to decrease, as shown in Figure 4.9B. Somewhat counterintuitively, increasing the probability of false positives from 0.0 to 0.1 could be considered to 'improve' the accuracy of isoform detection by shifting the distribution of the mean number of isoforms detected to slightly higher values and away from zero. This is probably because slightly increasing the probability of false positives allows some dropout events to be detected. In Figure 4.10, I investigate how the overlap fraction is affected by changes in the probability of false positives and negatives. I find that the overlap fraction increases as the probability of false positives increases, supporting the hypothesis that some dropout events are 'rescued' by false positive events. However, I note that in addition to 'rescuing' some dropouts, many unexpressed isoforms are also called as expressed, as indicated

by mean numbers of detected isoforms per gene per cell that are greater than one. When the probability of false positives and false negatives are equally increased (the diagonal of Figure 4.9A), the mean number of detected isoforms increases, suggesting that the increased rate of false positives dominates over the increased rate of false negatives. This is likely to be because more isoforms are unexpressed than are expressed, and thus there are more opportunities for false positive events than for false negative events. Overall, I find that high probabilities of false positives and false negatives decrease my ability to accurately detect expressed isoforms in scRNA-seq.

In Figure 4.9A, I showed that even when isoform quantification is 100% accurate, we underestimate the number of expressed isoforms for many genes. One hypothesis for why we are less able to detect isoforms in scRNA-seq data compared to in bulk RNA-seq data is that the sequencing depth is typically lower. A lower sequencing depth could mean that for many expressed isoforms, there are too few or no reads that would allow the expressed isoform to be uniquely identified.

To investigate whether sequencing depth could explain the difference in our ability to detect isoforms in bulk and scRNA-seq, I first identified a matched bulk and scRNA-seq dataset. The dataset I selected was a mouse Embryonic Stem Cell (mESC) dataset in which mESCs were cultured in 2i + LIF media (Kolodziejczyk et al., 2015). In the mESC dataset, each cell was sequenced to approximately 7 million reads on average, whilst the matched bulk data was sequenced to approximately 44 million reads.

To determine whether sequencing depth was responsible for the difference in our ability to detect isoforms in bulk and scRNA-seq, I randomly downsampled the bulk mESC RNA-seq dataset to 7 million reads 50 times. Using the orginal, undownsampled bulk RNA-seq dataset as the ground truth, in Figure 4.11 I plotted the mean overlap fractions for each gene in the downsampled bulk RNA-seq dataset and the matched scRNA-seq dataset. I found that the mean overlap fraction was significantly higher ($p < 2.2x10^{-}16$, Welch two sample t-test) for the downsampled bulk RNA-seq than for the matched scRNA-seq. This indicates that a lower sequencing depth does reduce our ability to detect isoforms, but that this does not fully explain the reduction in ability to detect isoforms between bulk and scRNA-seq. One

explanation for the reduction in ability to detect isoforms in scRNA-seq, over and above the reduction expected due to reduced sequencing depth, is that there could be heterogeneous isoform expression between individual cells. If this were the case, using the isoforms detected in bulk RNA-seq as the ground truth would not be appropriate. There are also potential technical explanations for the reduced ability to detect isoforms using scRNA-seq. For example, the enzymatic reactions associated with library preparation may have reduced efficiency when there is a lower amount of starting material, as is the case for scRNA-seq. Determining to what extent heterogeneous isoform expression and technical factors are responsible for our reduced ability to detect isoforms in scRNA-seq will require further study of cellular isoform heterogeneity and the technical noise associated with scRNA-seq.

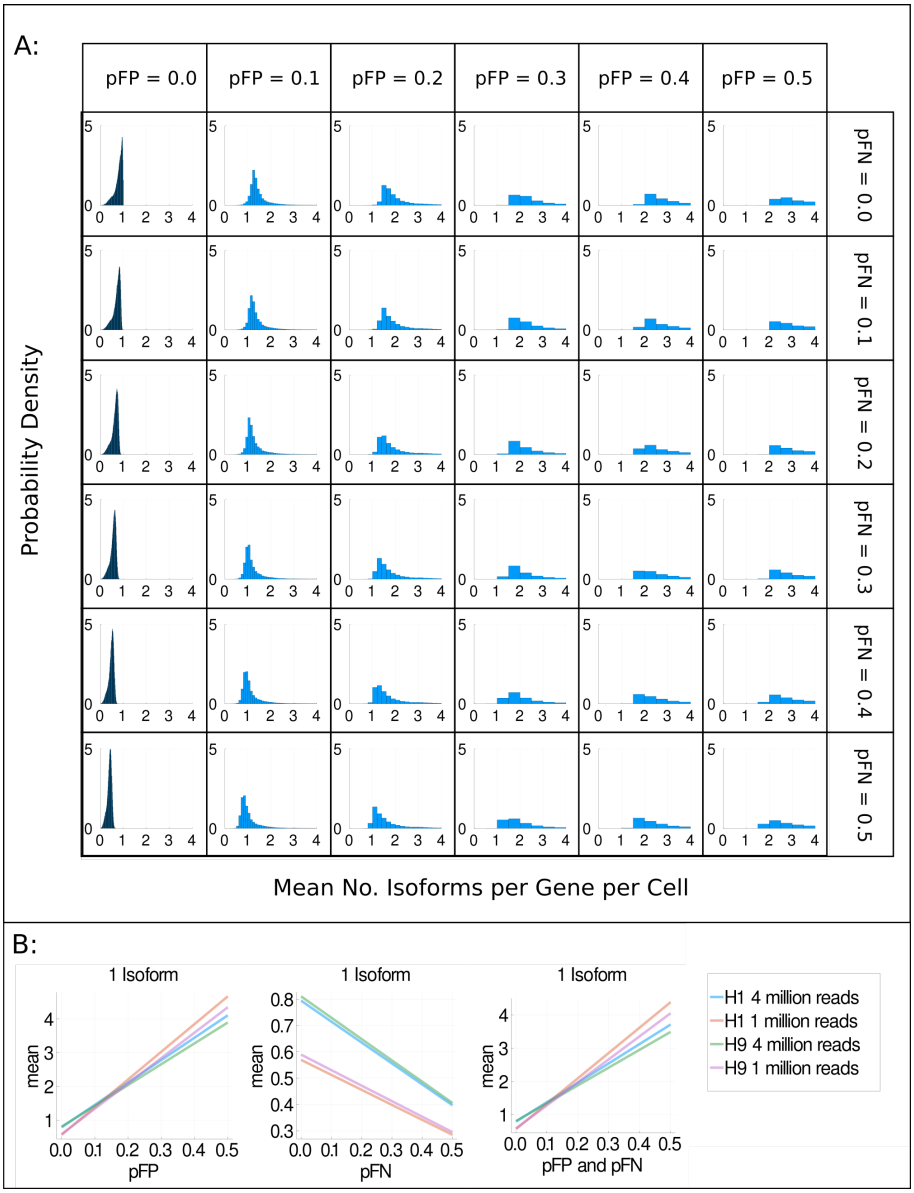


Figure 4.9: The impact of quantification errors on isoform detection. **A** Distributions of the mean number of isoforms detected per gene per cell when one isoform is expressed per gene per cell. The probability of false positives ($pFP$) increases from left to right and the probability of false negatives ($pFN$) increases from top to bottom. The dataset shown is H1 hESCs whose cDNA was split and sequenced at approximately 4 million reads per cell on average. **B** Summary plots of the average of the mean number of isoforms detected per gene per cell when $pFP$, $pFN$, or $pFP$ and $pFN$ are increased.

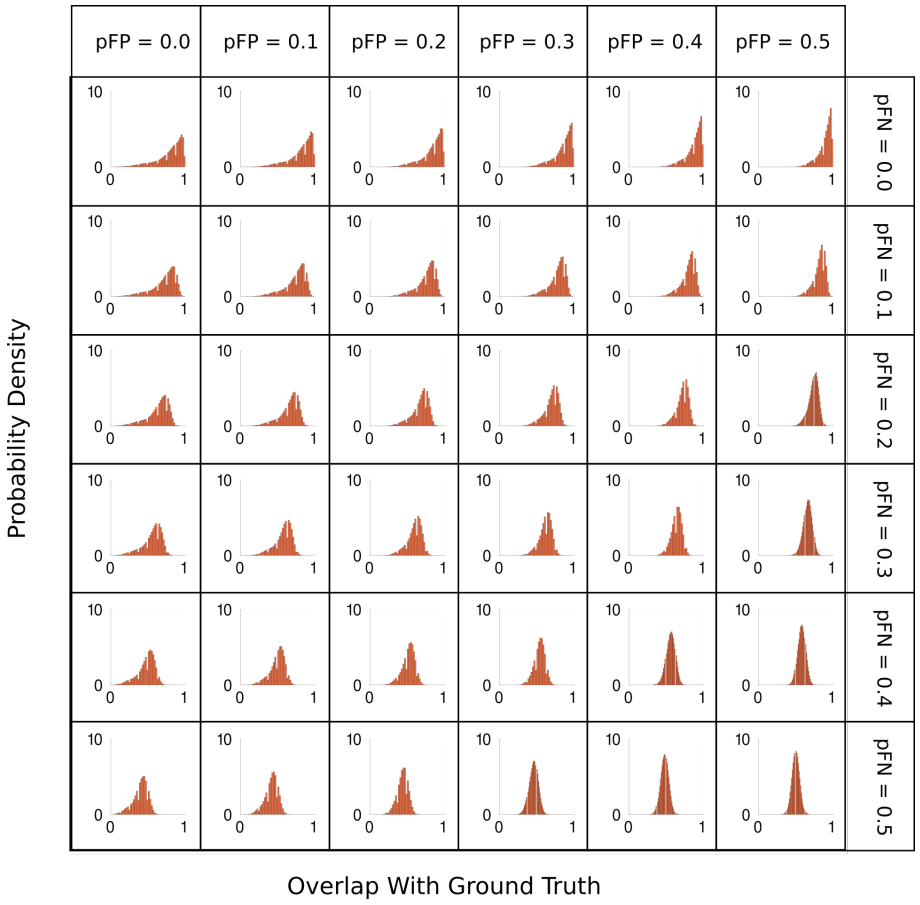


Figure 4.10: The impact of quantification errors on isoform detection. Distributions of the overlap fraction with the ground truth when one isoform is expressed per gene per cell. The probability of false positives ($pFP$) increases from left to right and the probability of false negatives ($pFN$) increases from top to bottom. The dataset shown is H1 hESCs whose cDNA was split and sequenced at approximately 4 million reads per cell on average.

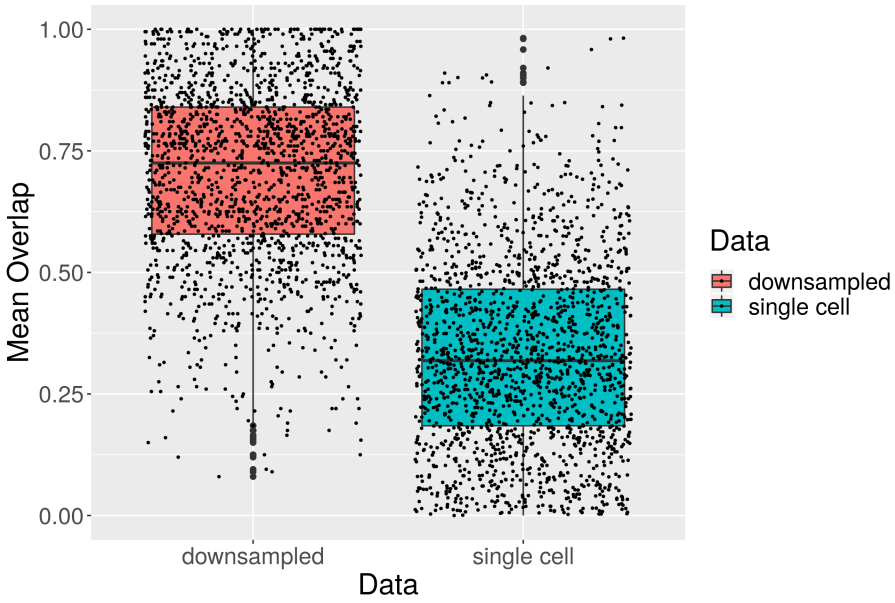


Figure 4.11: Boxplots of the mean overlap for each gene in the downsampled bulk and matched scRNA-seq datasets. The mean overlaps for each gene are overlaid on the boxplots as black points. Plots shown for Kolodziejczyk et al. mESCs cultured in standard 2i media + LIF (Kolodziejczyk et al., 2015).

### 4.1.2 Different models of isoform choice meaningfully change our simulation results

It is possible that different mechanisms of isoform choice at the cellular level could alter our ability to correctly detect which isoforms are present in scRNA-seq. Because there is uncertainty over the mechanism of isoform choice within single cells, I implement four different models of isoform choice in our simulations. I then ask whether different models of isoform choice alter the mean number of detected isoforms per gene per cell in our simulations.

I give a detailed description of how each of these models was implemented in the Methods chapter, here I provide a brief description of each model and the rationale behind it. I first model the alternative splicing process as a type III Weibull distribution, using a model described by Hu et al. (Hu et al., 2017). Based on observations about the molecular process of alternative splicing, Hu et al. suggested that the process could be well modelled by an extreme value distribution, and they found that a Weibull distribution best fits the expression levels of isoforms in bulk RNA-seq. In my second implemented model, I attempt to infer the probability of each isoform being 'chosen' to be expressed in a cell. I calculate the probability of an isoform being chosen based on the observed probability of the isoform being detected. My third model is identical to the second except that I allow the probability of an isoform being 'chosen' to vary between cells. I achieve this by sampling the probability of an isoform being chosen from a Beta distribution, using a similar approach as Velten et al. (Velten et al., 2015). In my final model, I choose a random number between 0 and 1 for each isoform. The random number is assigned to be that isoform's probability of being chosen, weighted against the probabilities of the gene's other isoforms being chosen. For brevity, I will refer to these four models as the Weibull model, the inferred probabilities model, the cell variability model and the random model below.

Figures 4.12 & 4.13 show the distributions of the mean number of detected isoforms when one, two, three or four isoforms are expressed per gene per cell for each model. Importantly, the distributions visibly differ between models. To quanti-

tatively confirm this, I perform a K-sample Anderson-Darling test on each row of graphs in Figures 4.12 & 4.13. I find that the distributions for 1, 2 and 3 isoforms significantly differ between the isoform choice models ($p<0.001$, see Appendix 3 Tables for details). In contrast, the distributions for 4 isoforms have a p-Value of 1.0 (1 million reads) and 0.999999 (4 million reads), consistent with these distributions originating from the same population. This is as expected, as in the 4 isoform simulations all of the isoforms are picked, and thus we would not expect isoform choice to matter. My qualitative and quantitative analyses indicate that different mechanisms of isoform choice alter my ability to detect splice isoforms in scRNA-seq. Therefore, a better understanding of the mechanism of isoform choice across the transcriptome could be key to enabling splicing analysis using scRNA-seq data. Without knowing how best to model isoform choice, my results suggest the presence of a substantial confounder.

My simulation results when using the inferred probability model compared with the cell variability model are almost identical. Given that the only difference between these models is whether or not isoform preference is allowed to vary between cells, this indicates that cellular heterogeneity in isoform preference does not change our ability to detect isoforms under the inferred probability model. I perform a K-sample Anderson-Darling test between the inferred probabilities and cell variability models for each row of Figures 4.12 & 4.13, and I find that these distributions do not significantly differ (see Appendix 3 Tables). I also observe that the results of the random model of isoform choice look more like the inferred probability and cell variability models than the Weibull model. This could be because the Weibull model determines the probability of an isoform being chosen based on the rank of that isoform, whereas all of the other models do not use a rank based approach. These observations and the difficulty I have interpreting them illustrate the need for a better understanding of how best to model isoform choice.

I hypothesise that the reason that different models of isoform choice differ in ability to detect isoforms could be because some models of isoform choice prefentially pick isoforms with a low probability of dropout, whereas other models do not exhibit this preference. To investigate whether different models of isoform choice differ in

their preference for picking isoforms with a low probability of dropout, in Figures 4.14-4.17, I plot the distributions of the probabilities of dropout for the isoforms chosen when one, two, three or four isoforms are picked using each of our four models. I would expect models with a preference for picking isoforms with a low probability of dropout to have distributions of dropout probabilities more skewed towards zero when small numbers of isoforms are chosen. When larger numbers of isoforms are chosen, I would expect to observe less skewed distributions, because the model is effectively forced to choose isoforms with higher probabilities of dropout due to a lack of alternatives. In contrast, if a model had no preference for picking isoforms with a low probability of dropouts, I would expect the distributions of the probabilities of dropout to be identical regardless of whether one, two, three or four isoforms are chosen.

In Figures 4.14-4.17, I find that only the Random model does not exhibit any preference for choosing isoforms with a low probability of dropout. Of the Weibull, inferred probability and cell variability models, the Weibull model has the dropout probability distribution most skewed towards zero when one isoform is picked, indicating that the Weibull model has the strongest preference for picking isoforms with a low probability of dropout. The Weibull model also detects the highest mean number of isoforms per gene per cell when one isoform is expressed in the ground truth, consistent with the hypothesis that the difference in the performance of the isoform choice models may be related to their preference for picking isoforms with a low probability of dropout.

If isoform detection ability of the isoform choice models is mainly determined by their preference for picking isoforms with a low probability of dropout, I would expect that if the probability of dropout was globally changed, it would alter the isoform choice models' abilities to detect isoforms. I investigate this in Supplementary Figure 4.18 by sampling dropout probabilities from the Beta distributions shown in Figure 4.7B. I find that more isoforms are detected by all isoform choice models when dropouts are sampled from distributions that are more skewed towards zero. This supports the hypothesis that choosing isoforms with a low probability of dropout improves the ability of isoform choice models to accurately detect isoforms.

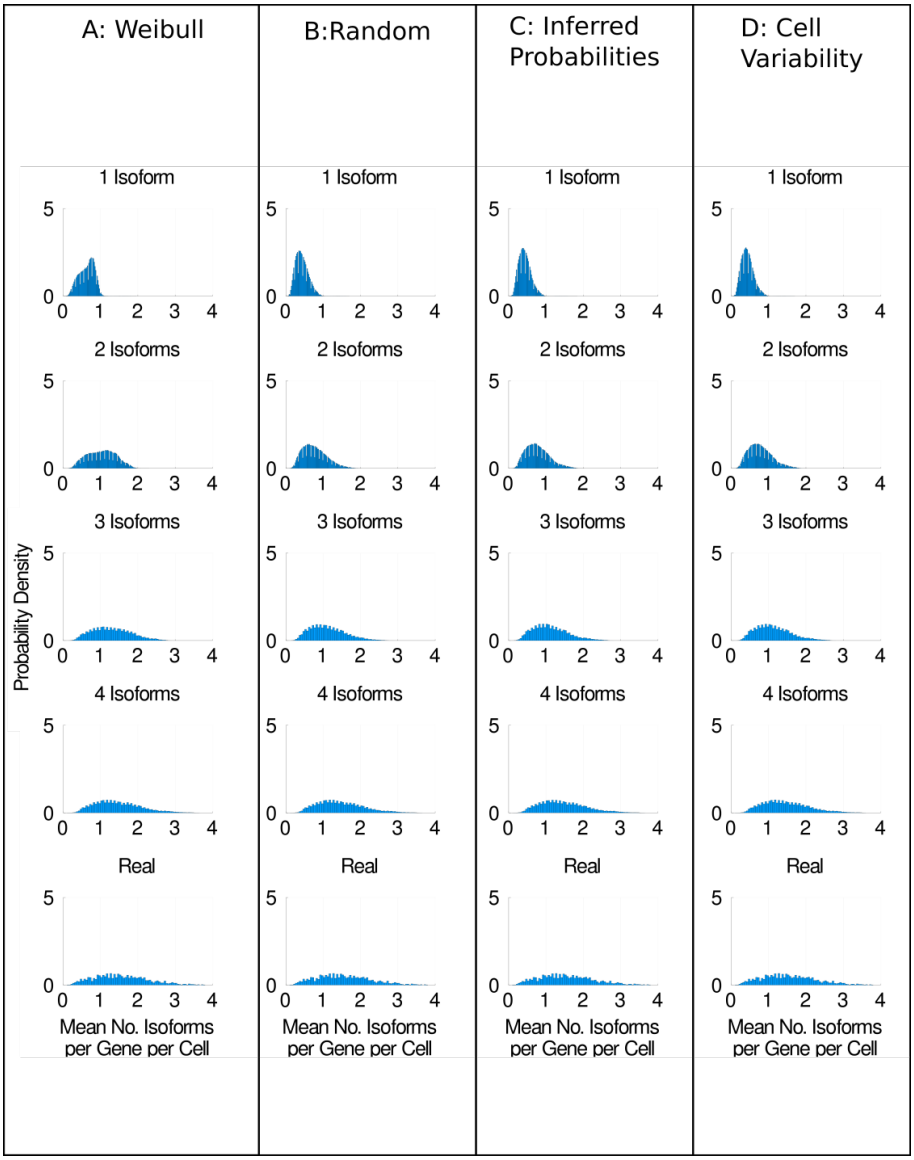


Figure 4.12: Different models of isoform choice alter our ability to detect isoforms. **A** Distributions of the mean number of isoforms detected per gene per cell for H1 hESCs sequenced at approximately 1 million reads per cell using the Weibull model of isoform choice (Bacher et al., 2017; Hu et al., 2017). **B** shows the same distributions when the random model is used. **C** shows the distributions when the inferred probabilities model is used. **D** shows the distributions when the cell variability model is used. See the main text for a detailed description of each model. Equivalent plots for the H9 datasets can be found in Appendix 3, Figures 9.3 & 9.5

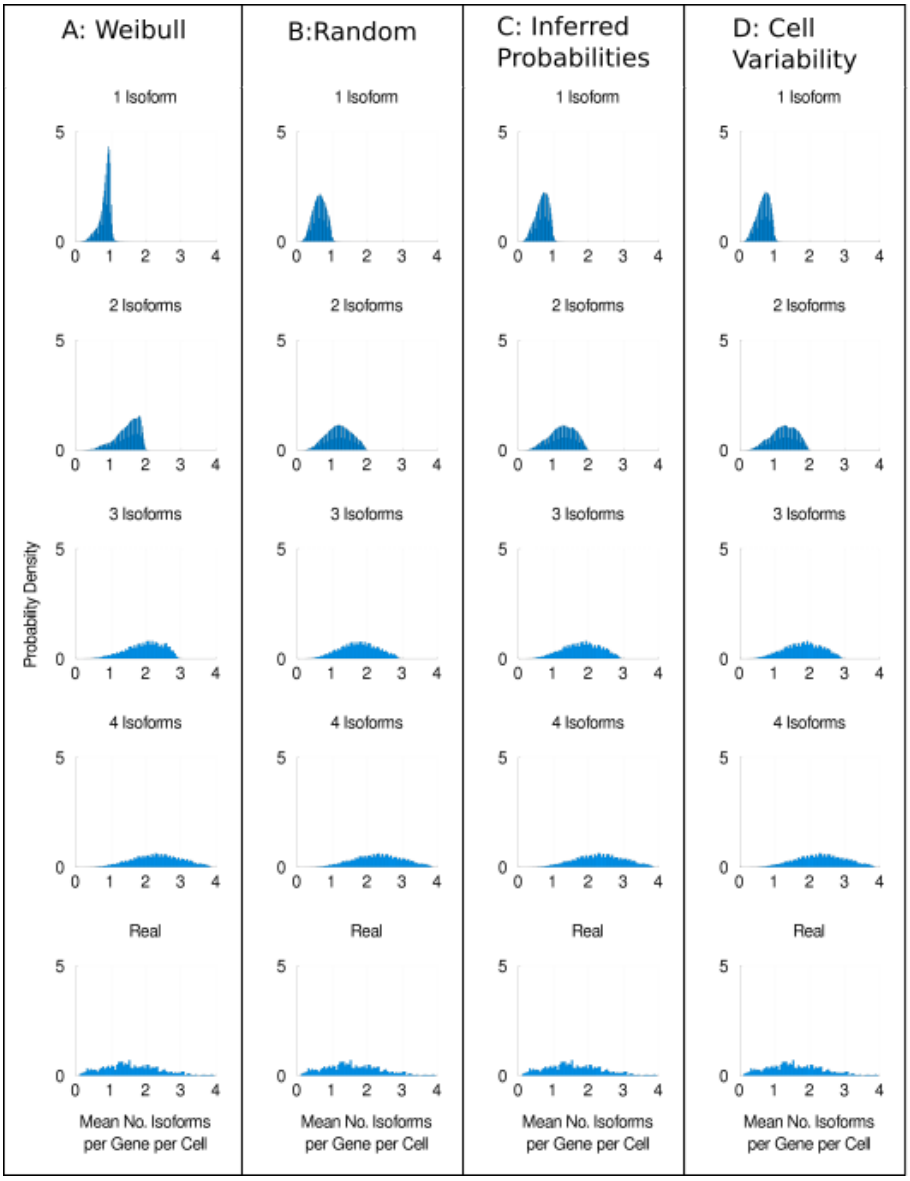


Figure 4.13: Different models of isoform choice alter our ability to detect isoforms. **A** Distributions of the mean number of isoforms detected per gene per cell for H1 hESCs sequenced at approximately 4 million reads per cell using the Weibull model of isoform choice. **B** shows the same distributions when the random model is used. **C** shows the distributions when the inferred probabilities model is used. **D** shows the distributions when the cell variability model is used. See the main text for a detailed description of each model. Equivalent plots for the H9 datasets can be found in Appendix 3, Figures 9.3 & 9.5

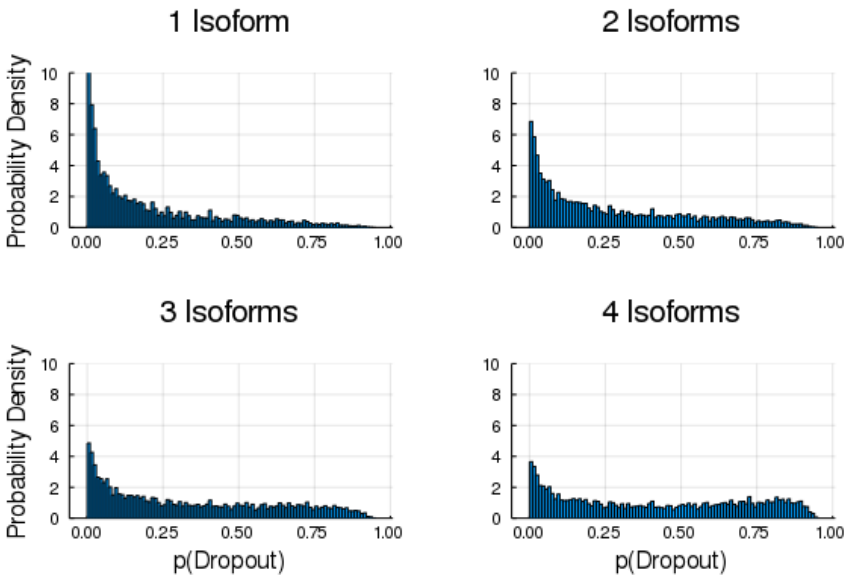


Figure 4.14: Distributions of the probabilities of dropouts for the isoforms selected by the Weibull model when one, two, three and four isoforms were picked by the model.

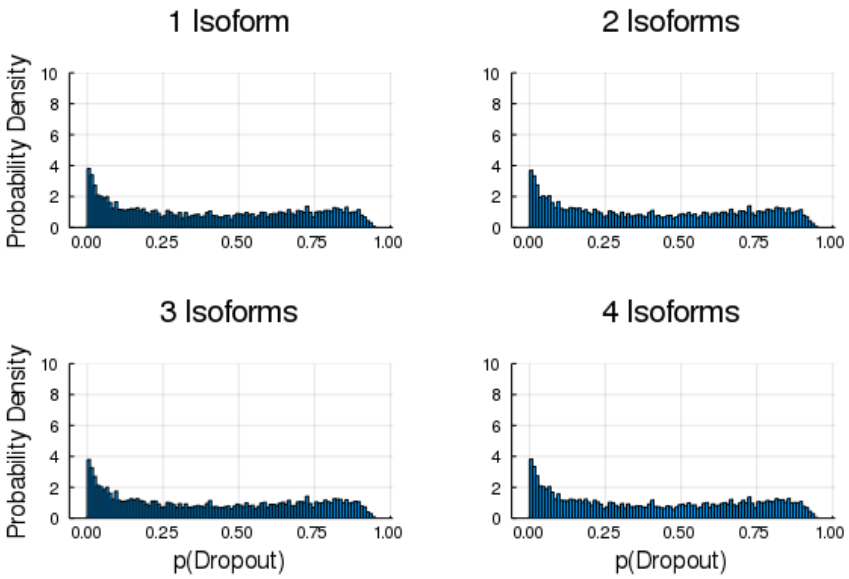


Figure 4.15: Distributions of the probabilities of dropouts for the isoforms selected by the Random model when one, two, three and four isoforms were picked by the model.

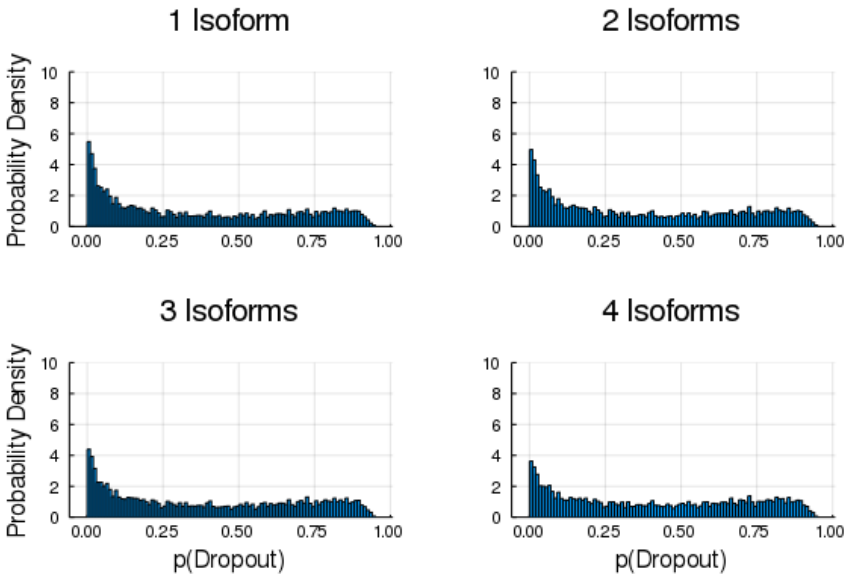


Figure 4.16: Distributions of the probabilities of dropouts for the isoforms selected by the inferred probabilities model when one, two, three and four isoforms were picked by the model.

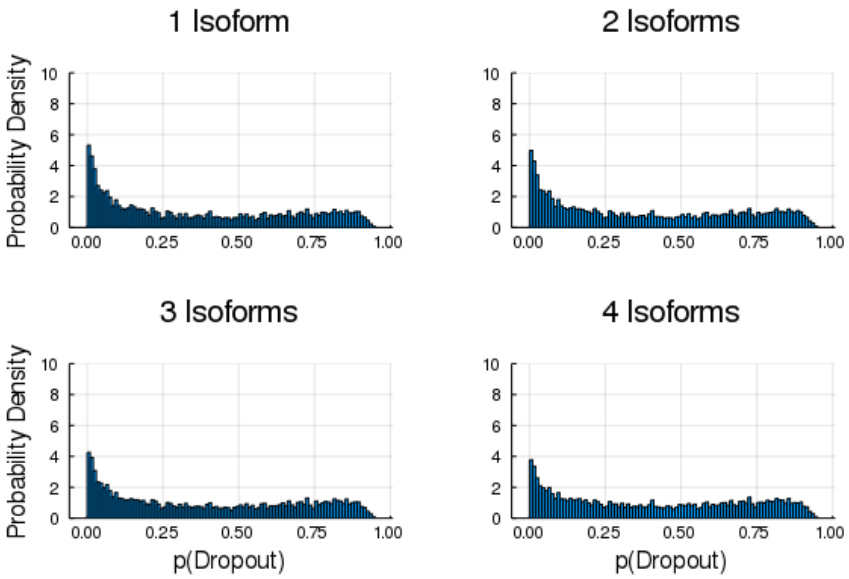


Figure 4.17: Distributions of the probabilities of dropouts for the isoforms selected by the cell variable model when one, two, three and four isoforms were picked by the model.

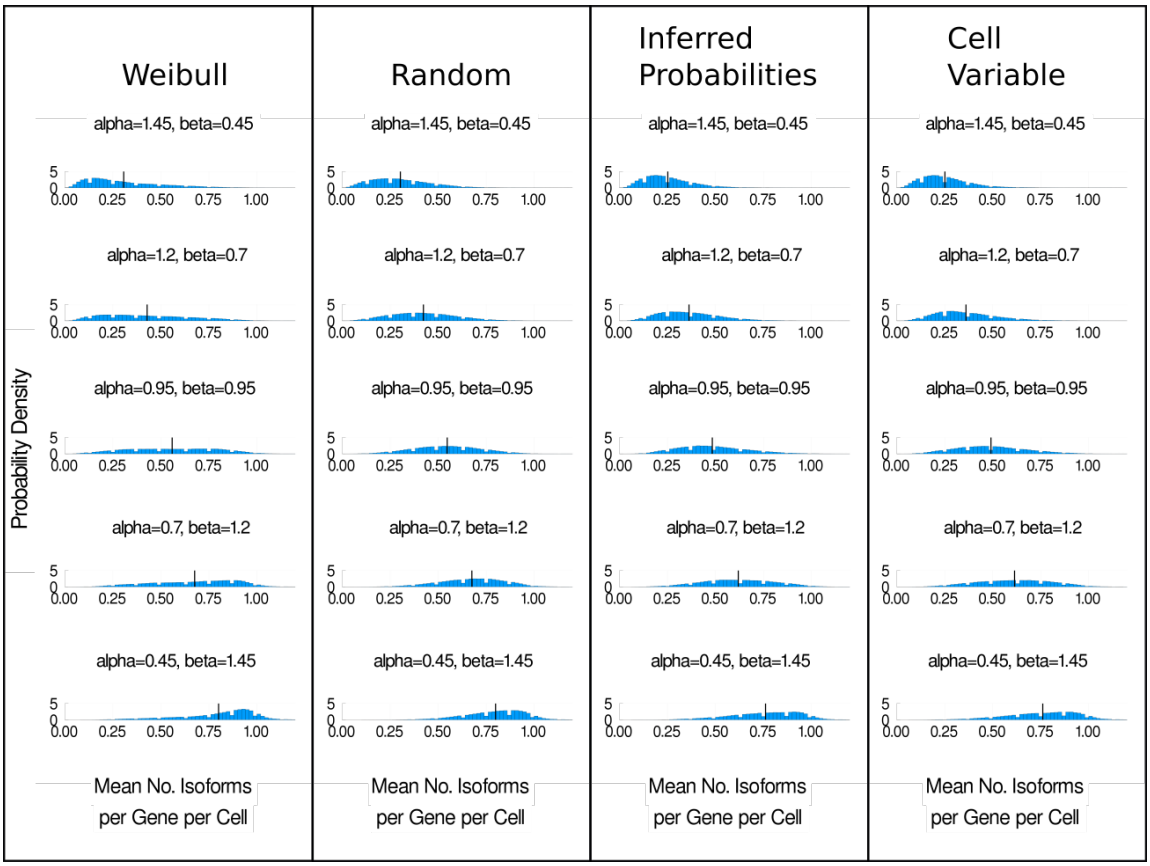


Figure 4.18: Distributions of the mean number of isoforms detected per gene per cell under different isoform choice models when dropout probabilities are sampled from the Beta distributions in Figure 4.7B.

### 4.1.3 Some models of isoform choice are more plausible than others

In the previous section, I observed that our simulation results for the inferred probability and cell variability models were extremely similar. To investigate how general my observation that allowing isoform preference to vary between cells does not alter our simulation results is, I developed three additional models of isoform choice. In the first model, the probability of selecting each isoform was sampled from a truncated Normal distribution with a mean of 0.25 and a standard deviation of 0.06 in each cell. In the second model, I sample the probability of selecting each isoform from a Bernoulli distribution, in which the value 1 is chosen 25% of the time and the value 0 is chosen 75% of the time in each cell. In the final model, the probability of selecting each isoform is always 0.25 (the 'p=0.25' model). The three models are illustrated in Figure 4.19A and additional details are given in the Methods chapter. Under the Normal and the Bernoulli models, the probability of picking each isoform varies between cells, whereas the probability of picking each isoform is constant between cells under the p=0.25 model. Importantly, although the distributions I am sampling isoforms from have very different shapes, the mean probability of picking each isoform is 0.25 for all three distributions.

In the second to fifth rows in Figure 4.19, I show the distribution of the mean number of isoforms detected per gene per cell when we simulate one isoform being expressed per gene per cell. There is no visible difference between my simulation results in each row regardless of which model of isoform choice is used. This is supported by a non-significant result in a K-sample Anderson-Darling test ($p = 0.998$). These findings are consistent with the hypothesis that my simulation results are unchanged whether or not the model of isoform choice used allows cell variability in isoform choice. I suggest that this is because we are reporting the mean number of isoforms detected per gene per cell in our simulations. Across many cells and rounds of simulation, the mean probability of selecting isoforms seems to determine the shape of our simulation result distributions, whereas the higher moments of the isoform choice probability distribution are apparently unimportant. Thus, includ-

ing cell variability in our isoform choice model appears not to matter. For future scRNA-seq studies in which the mean number of isoforms detected per gene per cell is an important metric, I conjecture that there is no need to model cellular variability in isoform choice, regardless of whether or not such variability exists in reality. Of course, if future studies are interested in precisely what isoforms are present in individual cells rather than a population mean, understanding whether or not cell variability in isoform choice exists is likely to be important.

I have established that our ability to detect isoforms using scRNA-seq is severely affected by the high rate of dropouts in scRNA-seq. Therefore, attempts to infer a biologically meaningful model of isoform choice from scRNA-seq data are likely to fail. However, I can make some general observations to help rule out certain models of isoform choice. In Figure 4.20A, I have ranked isoforms by their mean expression relative to other isoforms from the same gene (so for example, an isoform with rank 1 has the highest mean expression, an isoform with rank 2 has the second highest mean expression, and so on). Unsurprisingly, we find that the most highly ranked isoforms are substantially more highly expressed than lowly ranked isoforms. This is consistent with the finding that many genes appear to have a 'major', more highly expressed isoform, and one or more 'minor', less highly expressed isoform (Wang et al., 2008; Gonzàlez-Porta et al., 2013). I suggest that this behaviour needs to be represented in some way in future models of isoform choice, and models that do not represent it (for example, our Random, Normal, Bernoulli and p=0.25 models) are probably overly simplistic. In Figure 4.20B I rank isoforms by their probability of dropout, where the isoform with the lowest probability of dropout compared to other isoforms from the same gene has rank 1. I observe a very similar pattern in which highly ranked isoforms have a substantially lower probability of dropout relative to lowly ranked isoforms, further supporting the finding that 'major' and 'minor' isoforms exist for many genes. I find a similar pattern of results for the H9 hESCs and the H1 hESCs sequenced at 4 million reads, shown in Figures 4.21 - 4.23.

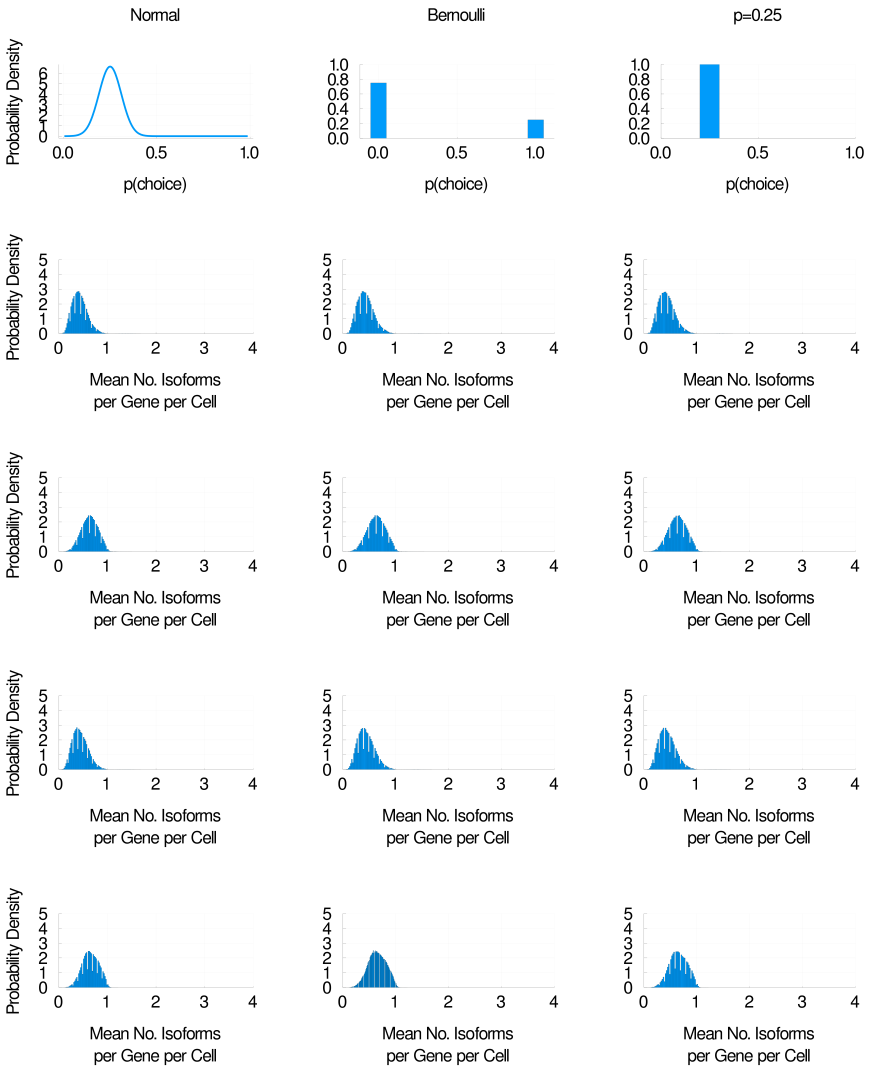


Figure 4.19: Some models of isoform choice are more plausible than others. We model the probability of picking any given isoform as a Normal distribution, a Bernoulli distribution and a constant probability, all with the same mean (0.25) (top row of graphs). In the following rows, I show the distributions of the mean number of isoforms per gene per cell detected when each model of isoform choice is used. The second row is H1 hESCs sequenced at 1 million reads, the third row is H1 hESCs sequenced at 4 million reads, the fourth row is H9 hESCs sequenced at 1 million reads and the fifth row is H9 hESCs sequenced at 4 million reads.

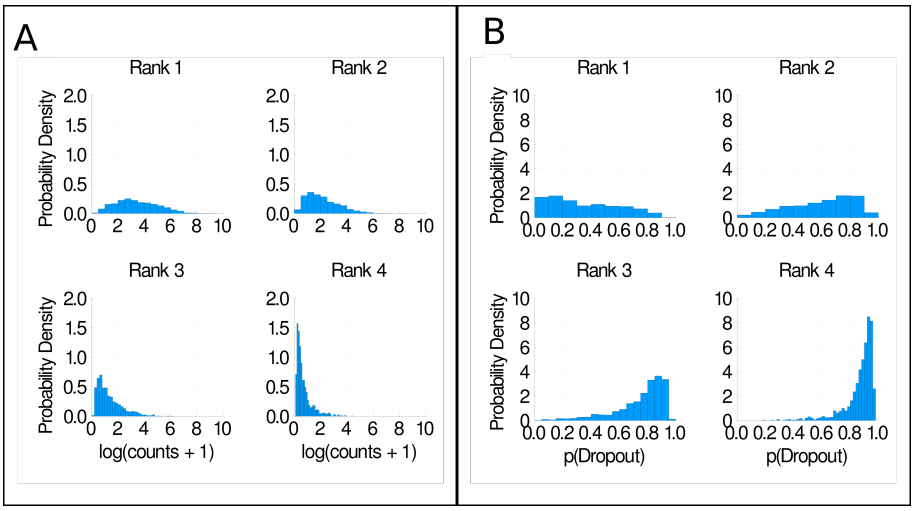


Figure 4.20: **A** Histograms of mean isoform expression, ordered by isoform rank. **B** Histograms of dropout probability, ordered by isoform rank. All plots shown are for H1 hESCs sequenced at 1 million reads per cell.

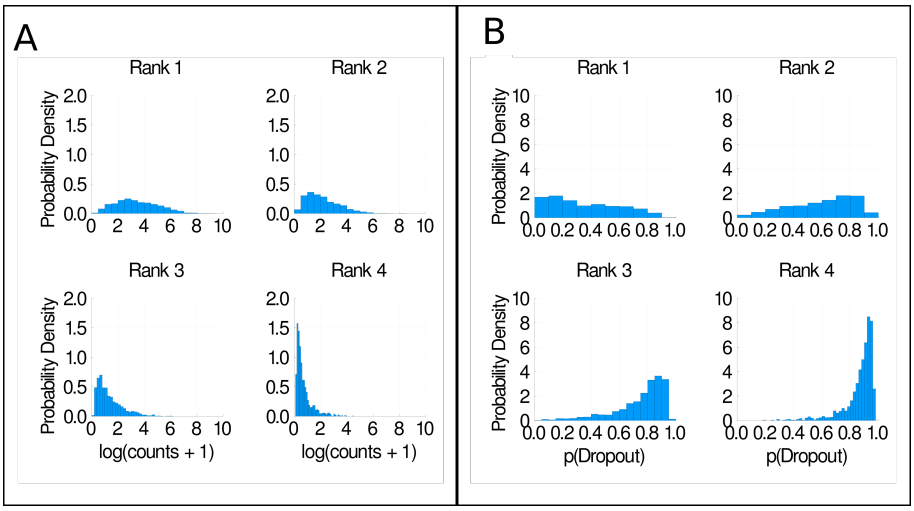


Figure 4.21: **A** Histograms of mean isoform expression, ordered by isoform rank. **B** Histograms of dropout probability, ordered by isoform rank. All plots shown are for H1 hESCs sequenced at 4 million reads per cell.

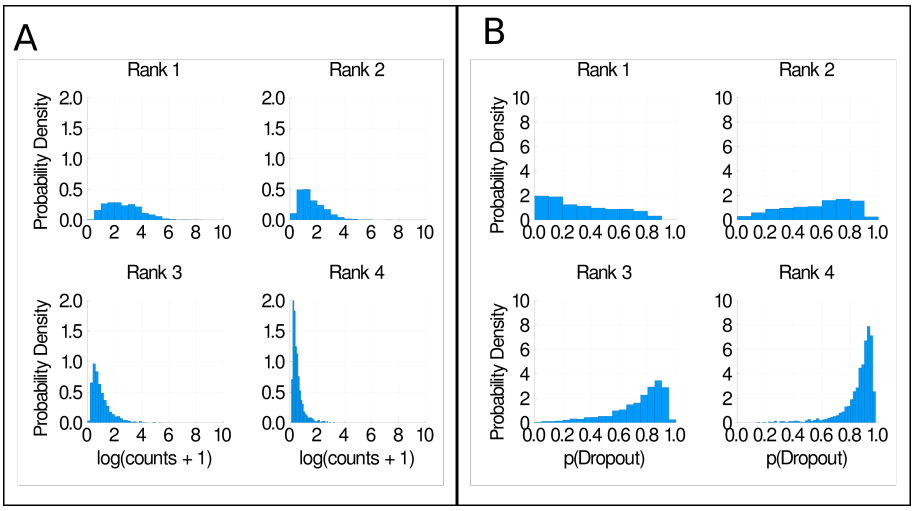


Figure 4.22: **A** Histograms of mean isoform expression, ordered by isoform rank. **B** Histograms of dropout probability, ordered by isoform rank. All plots shown are for H9 hESCs sequenced at 1 million reads per cell.

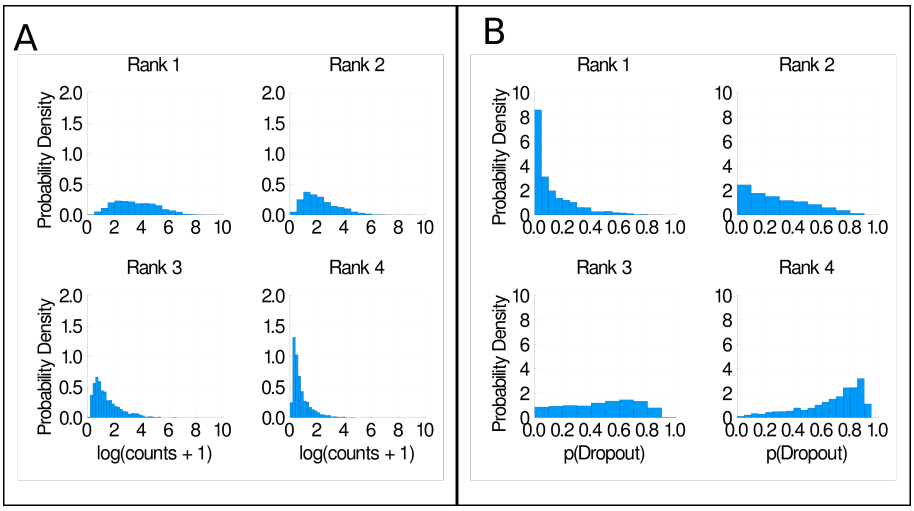


Figure 4.23: **A** Histograms of mean isoform expression, ordered by isoform rank. **B** Histograms of dropout probability, ordered by isoform rank. All plots shown are for H9 hESCs sequenced at 4 million reads per cell.

### 4.1.4 A mixture modelling approach suggests genes for which four isoforms are detected typically express around three isoforms per cell

I ask whether my simulation based approach could shed any light on the biological question of how many isoforms are expressed per gene per cell. To do this, I simulate one, two, three and four isoforms being expressed per gene per cell and compare the mean isoforms detected distributions to the distribution of isoforms detected per gene per cell for genes for which four isoforms were detected in the real dataset (see Figure 4.24A & B). I then approximate each distribution as a log normal distribution and take a mixture modelling approach to estimate the mixing fraction for each of our simulated distributions in the real distribution.

Figure 4.24C shows the mixing fractions found over 100 iterations of expectation maximisation for H1 hESCs sequenced at approximately 1 million reads per cell. In Figure 4.24C, the mixing fraction for the distribution corresponding to four isoforms being expressed per gene per cell is over 90%. This suggests that genes detected to express four isoforms in this dataset typically express four isoforms per gene per cell. However, in Figure 4.24D, after 100 iterations of expectation maximisation for H1 hESCs sequenced at 4 million reads per cell, the distribution with the largest mixing fraction is that corresponding to three isoforms per gene per cell. This suggests that genes detected to express four isoforms in this dataset most often express three isoforms per gene per cell. As the cDNA sequenced at 1 and 4 million reads per cell came from the same population of cells, it is unlikely that both of these statements are true. I propose several possible explanations for why we might observe this result.

First, I might be over-estimating the dropout rate at 1 million reads per cell. As there is less information with which to infer the dropout rate at 1 million reads per cell compared to at 4 million reads per cell, it is plausible that our estimates of the dropout rate are less accurate at 1 million reads per cell. Whether or not there is a systematic bias towards over-estimating the dropout rate at low sequencing depths is unknown, and goes beyond the scope of this thesis.

Second, I have established that the model of isoform choice influences the outcome

of our simulations but we do not know which model of isoform choice is correct. Therefore I am (almost certainly) attempting to fit distributions that do not represent reality. Figure 4.24 shows my mixture modelling approach using the Weibull model of isoform choice. I note however that fitting our alternative models of isoform choice achieves a similar result, in that the largest mixing fraction goes to four isoforms at 1 million reads per cell and to three isoforms or fewer at 4 million reads per cell (see Appendix 3 Figures 9.8-9.14).

Third, the genes detected to express four isoforms differ between the sequencing depths of 1 and 4 million reads. More genes are detected to express four isoforms at 4 million reads (1443 versus 1543 for the H1 cells, 1453 versus 1524 for the H9 cells). Whilst this is not a dramatic difference, it does mean that the mixing fractions between these two depths could genuinely differ, although this is unlikely to fully explain the observed difference.

Fourth, I assume all genes for which four isoforms are detected in the real data actually express four isoforms. Due to dropouts and quantification errors, this may not be accurate, and some genes for which four isoforms are detected may express a different number of isoforms in reality.

Fifth, my parameter estimation for quantification errors and isoform choice modelling is not one hundred percent accurate. I cannot rule out that this could be confounding the results of our mixture modelling approach.

My mixture modelling experiments broadly support the hypothesis that it might be common for a cell to produce more than one isoform per gene. However, there are clearly a lot of potential confounders in our approach, many of which relate to uncertainty about dropouts, quantification errors and isoform choice. I note that without having either a ground truth knowledge of how many isoforms are produced from given genes in given cells, or good estimates of dropout probabilities, quantification errors and isoform choice mechanism, it is hard to imagine how an accurate and reliable estimate of the number of isoforms produced per gene per cell could be obtained.

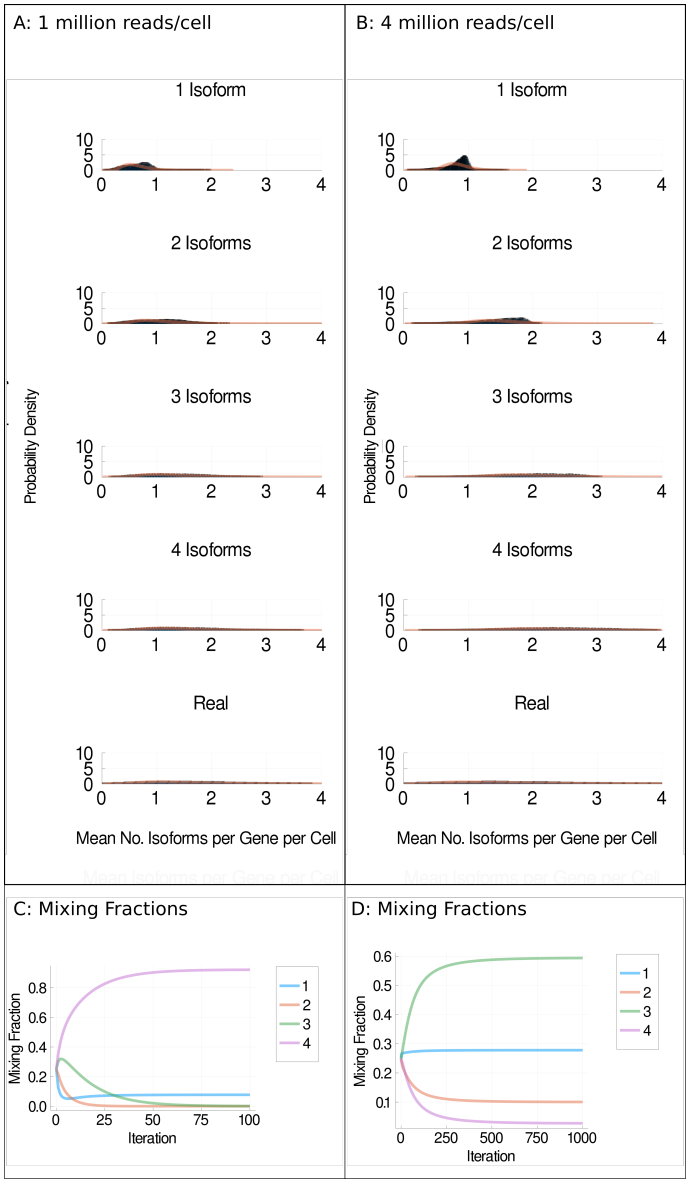


Figure 4.24: Mixture models. **A** and **B** Distributions of detected isoforms per gene per cell (blue) and log normal fitted distributions (orange) for H1 cells sequenced at 1 million reads per cell (**A**) or 4 million reads per cell (**B**) under the Weibull model. **C** and **D** Mixing fractions vs iterations of expectation maximisation for 1 million reads per cell (**C**) and 4 million reads per cell (**D**). Each coloured line represents the distributions for one, two, three or four isoforms being simulated as expressed per gene per cell. Equivalent plots for other isoform choice models and H9 cells can be found in Appendix 3, Figures 9.8-9.14.

## 4.2 Discussion

In this study, I use a novel simulation based approach to ask whether it is possible to study alternative splicing at the level of individual cells using scRNA-seq. In my simulations, I simulate four scenarios in which every gene produces one, two, three or four isoforms per gene per cell. That it is difficult to clearly distinguish between these four situations emphasises the challenges associated with distinguishing the much subtler and more complex patterns of alternative splicing that likely exist in reality. Whilst scRNA-seq is capable of detecting some splicing events, the confounding effect of dropouts means we are likely to underestimate the number of splicing events occurring in individual cells.

I next ask what limitations must be overcome to make alternative splicing analysis possible using scRNA-seq. I find that reducing the probability of dropouts improves our ability to accurately detect isoform number. Therefore, reducing the frequency of dropouts could be one method to improve the accuracy of splicing analyses in scRNA-seq. To some extent, this could be achieved by sequencing cells more deeply, although I note that at 4 million reads per cell we still substantially underestimate isoform number in the H1 hESCs. Unfortunately, extremely deeply sequenced datasets (eg. >10 million reads per cell) are likely to suffer more with PCR artefacts and potentially a higher false positive rate of isoform detection (Islam et al., 2014; Kanagawa, 2003). Fundamentally, the low capture efficiency of scRNA-seq is likely to be a consequence of a small amount of starting material. This can probably be rescued to some extent by more PCR cycles and sequencing at higher depths, however I would not expect this to fully solve the problem.

A more radical way to overcome confounders due to dropouts would be if scRNA-seq technologies changed in some fundamental way that increased capture efficiency. Whether this is feasible is unclear. Alternatively, I note that if we could estimate the probability of dropout for each isoform more accurately, in theory it should be possible to correct for confounding effects due to dropouts in splicing analyses. Therefore, to enable splicing analysis using scRNA-seq, either the capture efficiency of the technology needs to improve, or more work characterising the probability of

dropouts at an isoform level is required.

In this chapter, I exclusively considered the impact of technical dropouts on isoform detection. However, it is known that many genes are heterogeneously expressed, whether due to 'bursty' transcription or cell type specific expression (Urban and Johnston, 2018). Ideally, the impact of biological dropouts on isoform detection would be evaluated alongside the impact of technical dropouts. Unfortunately, to the best of my knowledge, there is currently no reliable methodology to distinguish between biological and technical dropouts. The goal of imputation approaches is to identify and correct for technical dropouts, but a recent benchmark found that imputation approaches often introduce a high rate of false positive results (Andrews and Hemberg, 2018b). This indicates that the problem of distinguishing between biological and technical dropouts is not yet solved. As it is not currently possible to resolve between biological and technical dropouts, it is also challenging to accurately model biological dropouts, as little is known about their prevalence and how the frequency of biological dropouts might vary with genomic features. I hope that future work in this space will enable more accurate identification of biological and technical dropouts, thus enabling studies such as mine to be extended to account for biological as well as technical dropouts.

Long read technologies could in theory enable 100% accurate isoform quantification, if issues due to a high base calling error rate could be overcome (Fu et al., 2019). However I find that even when no isoform detection errors occur, our ability to accurately detect isoforms is very limited. Therefore, long read technologies or isoform quantification software improvements alone are not sufficient to enable accurate splicing analysis in scRNA-seq. In addition, I note that at present, the read throughput of long read platforms is too low to enable meaningful isoform detection and quantification across a large number of cells (Arzalluz-Luque and Conesa, 2018). A more immediate way in which long read technologies could improve isoform quantification accuracy is by using long read technologies to improve transcriptome annotations. In many non-model organisms, a high proportion of isoforms are missing from reference transcriptomes, making the problem of isoform detection and quantification substantially harder. Long read approaches combined with tissue specific transcrip-

tome curation could dramatically improve isoform quantification accuracy in poorly annotated organisms. More accurate isoform detection and quantification would in turn improve our ability to gain biological insight from sequencing data collected from these organisms.

A limitation of the methodology in this chapter is that my approach for simulating quantification errors is very simplistic. In particular, I assume that the probability of a false positive or a false negative event is constant, and does not depend on the GC content, length, magnitude of expression or any other relevant features of the isoform being simulated. In reality, the probability of isoform detection errors probably does depend on factors such as GC content and how highly expressed the isoform is. However, relatively little research has been done into the relationship between features of isoforms, such as GC content and magnitude of expression, and the probability of isoform detection errors. Further research into how genomic and other features of isoforms affect the likelihood of isoform detection and quantification errors would enable more accurate error models to be built in future. This would be valuable both in studies such as this one and more generally, as it would enable more sophisticated error correction models to be developed.

Little is known about the biological process of isoform choice in individual cells for most genes. Thus, accurately modelling this process is challenging. I find that different models of isoform choice alter our simulation results. This indicates that without better understanding of the process of isoform choice, alternative splicing analyses are potentially confounded by this unknown factor. Research into the process of isoform choice within individual cells across the transcriptome would enable more accurate models of isoform choice to be built, reducing or removing this confounder from future alternative splicing analyses. An important finding from my study is that the ability of isoform choice models to accurately detect isoforms is correlated with the preference of isoform choice models for choosing isoforms with a low probability of dropout. It would therefore be highly relevant to establish whether cells have a preference for expressing isoforms with a low probability of dropout. Isoforms with a low probability of dropout are in practice usually isoforms which are highly expressed. Therefore, if cells have a preference for expressing highly expressed

isoforms with a low probability of dropout, I would expect it to be relatively easy to accurately detect how many isoforms are expressed in individual cells. In contrast, if it is common for cells to express lowly expressed isoforms with a high probability of dropout, I would expect it to be much harder to accurately detect the number of expressed isoforms using scRNA-seq. Establishing which scenario is more biologically relevant would therefore be highly valuable to the single cell community.

It is important to note that the probabilistic models of isoform choice used in our study are unlikely to be realistic models of isoform choice for two reasons. Firstly, we know little about the underlying biological process of isoform choice for most genes. Therefore at best the models we have devised in this study are educated guesses as to what the true underlying process might be. Secondly, it is likely that the isoforms chosen by our isoform choice models will have an impact on the probability of a quantification error occurring. Different isoforms have different read generation biases, and will generate reads with different mapping properties. In our simulations, we have not modelled the impact of, for example, different splice junction abundances on our ability to detect isoforms, although factors such as this are likely to have an impact on our ability to detect isoforms. I would welcome future studies addressing the more nuanced issues associated with the interplay between isoform choice and quantification errors, although I believe that a better understanding of how to accurately model isoform choice and quantification errors would be a prerequisite to such studies. If isoform expression is found to be heterogeneous between cells, interplay between isoform choice and isoform quantification errors could partly explain why we were less able to detect isoforms present in mESC scRNA-seq data than in downsampled bulk RNA-seq.

I am able to detect evidence in support of 'major' and 'minor' isoforms, and propose that future models of isoform choice should attempt to capture this behaviour. However, I note that whilst my observations help discard models of isoform choice, I believe that scRNA-seq is currently too confounded by dropouts to accurately infer a model of isoform choice at the single cell level. I suggest that smFISH would be a more appropriate technology to investigate how isoform choice is regulated in individual cells. Indeed, smFISH has previously been used to study alternative splicing

and isoform choice in individuals cells for a small number of genes (Velten et al., 2015; Ciolli Mattioli et al., 2019; Waks et al., 2011)

The results of our mixture modelling experiments are consistent with multiple isoforms being produced per gene per cell, however I note that our mixture modelling experiments are heavily confounded by a lack of understanding about dropouts, isoform choice and perhaps quantification errors to a lesser extent. Therefore, I argue that at this time, scRNA-seq will not be able to provide the answer to basic biological questions about how many isoforms are produced per gene per cell.

In addition to detecting isoforms, isoform quantification tools attempt to determine how highly expressed isoforms are. Isoform quantification is a substantially harder problem than isoform detection. Due to uncertainties over how highly expressed isoforms are in individual cells, how best to model PCR amplification bias and differences in library sizes between individual cells and how best to incorporate relative expression into a model of isoform quantification errors, I suspect isoform quantification is also likely to be substantially harder to model than isoform detection. For these reasons, I have focused on isoform detection in this study, but suggest that future work investigating our ability to detect the relative expression of isoforms would be highly valuable to the field. I note that although we have not directly evaluated our ability to resolve the relative expression magnitude of isoforms in this study, that we often struggle to accurately detect isoforms implies that we would often struggle to determine how highly expressed they are.

Based on my findings, at this time I do not recommend attempting alternative splicing analysis using scRNA-seq. As my analysis suggests that one of the greatest confounders in studying splicing is dropouts, it may be relatively safe to study alternative splicing using only highly expressed isoforms with very few dropouts. For many genes in many datasets, this severely limits the scope in which splicing can be studied. However, I make actionable suggestions for how splicing analysis could be enabled in the future. An improved understanding of the prevalence of technical dropouts at the isoform level could enable us to reduce confounding effects due to dropouts. Improvements to the capture efficiency of scRNA-seq would similarly reduce confounding effects due to dropouts. Increased study of isoform choice at

the single cell level using technologies such as smFISH would enable better models of isoform choice to be generated, eliminating confounders. Although I find quantification errors to be a relatively small confounder, further reducing quantification errors using long read technologies and more accurate quantification tools would be welcome. Although I have concluded that accurate alternative splicing analysis with scRNA-seq is not possible today, I am optimistic that it could become possible in the near future.

# Conclusions

At present, alternative splicing analyses using scRNA-seq are substantially confounded. Better characterisation of dropouts or improvements in capture efficiency would reduce confounding effects due to dropouts. Further research into the process of isoform choice at a single cell level would reduce confounding effects caused by a lack of knowledge about isoform choice. Quantification errors are a relatively minor confounder, although improvements in this area are still welcome. At present, to the best of my knowledge, a large scale unconfounded analysis of the number of isoforms produced per gene per cell has not been performed. Therefore, we still do not know how many isoforms are typically produced per gene per cell.

# 5

# Methods

*The proper method for inquiring after the properties of things is to deduce them from experiments.*

– Isaac Newton, quoted by (Strong, 1951)

## 5.1 Simulation based benchmarking of isoform quantification using scRNA-seq

The methods in section 5.1 are for the experiments carried out in chapter 2, and the experiments presented in Figures 3.1 & Figures 3.2.

### 5.1.1 Software tools

A variety of software tools were used in chapter 2. An overview of each tool and its assumptions are provided below. Table 5.1 is a summary table of the isoform quantification tools evaluated in this benchmark provided at the end of this subsection.

#### STAR

STAR was the aligner used in the RSEM simulations, and by RSEM and eXpress in the benchmark of isoform quantification (Dobin et al., 2013; Li and Dewey, 2011;

Roberts and Pachter, 2013). STAR was chosen as the aligner of choice in part because the algorithm underlying STAR is designed to facilitate the alignment of reads originating from alternatively spliced isoforms (Dobin et al., 2013). STAR's alignment algorithm works by first finding the longest genomic region that maps to the start of the read. If the read contains a single exon, the entire read is aligned in this step. However, many reads contain multiple exons which are separated by intronic sequences in the genome. If the read being aligned contains multiple exons, after finding the longest genomic region that maps to the start of the read, the algorithm proceeds to try and find the longest genomic region that maps to the unmapped portion of the read. This process is repeated until the entire read is aligned, potentially with some mismatches. STAR then stitches together the aligned regions of the read, initially searching for a complete alignment within a user-defined genomic window but also allowing for chimeric alignments. A user-defined score for matches, mismatches, indels and splice junction gaps, is used to identify the alignment with the lowest score. This alignment is the one reported by STAR. For multimapping reads, STAR reports all alignments within a certain score threshold of the optimal alignment. The main assumptions made by STAR are that the genomic window and penalty scores defined by the user are appropriate. In this benchmark, I used the default windows and scores provided by STAR, as I believe this is the most common user behaviour.

**RSEM**

RSEM is an isoform quantification tool and a read simulator, which was used both to simulate reads and to perform isoform quantification in my benchmark (Li and Dewey, 2011). To perform isoform quantification, RSEM first generates a set of reference transcript sequences. Importantly, RSEM aligns reads to these reference transcript sequences rather than the entire genome. Next, RSEM uses an aligner to align reads to these reference sequences. In my benchmark, STAR was used as the aligner (Dobin et al., 2013). Following alignment, RSEM uses an Expectation-Maximisation (EM) algorithm to estimate isoform abundance. RSEM's directed

graphical model is shown in Figure 5.1.

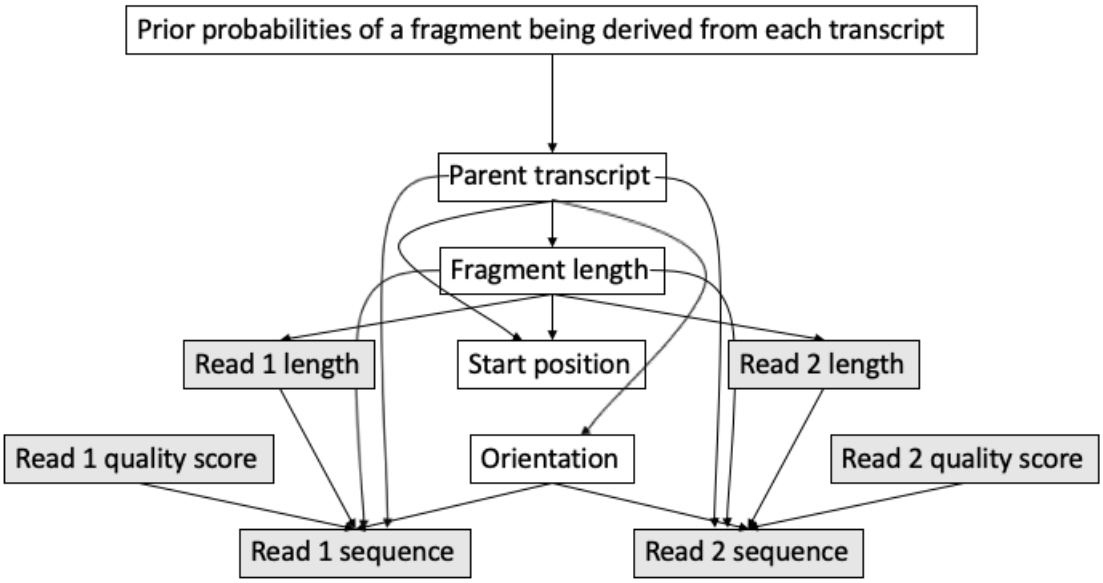


Figure 5.1: RSEM's directed graphical model. Observed variables, which can be directly observed from the reads data, are shown in grey. Latent variables, which must be inferred by the EM algorithm, are shown in white. This figure is adapted from Figure 4 in (Li and Dewey, 2011).

The goal of Expectation-Maximisation algorithms is to find the maximum likelihood parameters for equations which can not be directly solved. In the directed graphical model above, we know the value of the observed values in white boxes, but not of the unobserved or latent values in grey boxes. We infer the values of these latent variables using RSEM's EM algorithm. For the first twenty iterations of RSEM's EM algorithm, and for every hundredth iteration after, the values of the orientation, read start position distribution, fragment length and parent transcript are updated. In all other iterations, RSEM only updates the prior probabilities of a fragment being derived from each transcript. The algorithm continues until the prior probabilities of a fragment being derived from each transcript converge (ie. the

prior probabilities no longer change meaningfully between one EM iteration and the next).

Once RSEM has completed quantification, in addition to outputting expression estimates, it saves the values it inferred for the latent variables in its graphical model. This enables RSEM to act as a reads simulator, using the latent variable parameters it inferred from real data. When RSEM uses its graphical model to probabilistically simulate reads, it counts where each read originated in the transcriptome. Thus RSEM can be used to generate reads data with ground truth expression data.

The core assumption of RSEM is that its directed graphical model combined with the EM algorithm is capable of generating accurate estimates of isoform expression, meaning that underlying asumptions about the fragment length distribution and relationships between the variables illustrated in Figure 5.1 must be at least somewhat realistic. When RSEM simulates reads, a key assumption in the ground truth expression is that every expressed transcript generates one and only one read. This is unlikely to be true, especially for scRNA-seq. Therefore, benchmarks of isoform quantification tools using RSEM as a read simulator assess the ability of quantification tools to correctly determine where reads originated from in the transcriptome, as oppose to their ability to determine what transcripts were originally present in the cell(s).

### eXpress

eXpress is an isoform quantification tool whose performance was evaluated in my benchmark (Roberts and Pachter, 2013). eXpress takes a bam file of aligned reads as input. In my benchmark, STAR produced the input bam file (Dobin et al., 2013).

eXpress uses an online algorithm, meaning that eXpress's algorithm does not require the entire input all at once at the start, but can process the input piece by piece in the order it arrives. The authors suggest this could enable eXpress to be coupled directly to a sequencer that produced reads one at a time. Whether this was ever realised is unclear.

Like RSEM, eXpress uses a directed graphical model combined with an EM al-

gorithm to generate its expression estimates (Li and Dewey, 2011). eXpress's EM algorithm is illustrated in Figure 5.2.

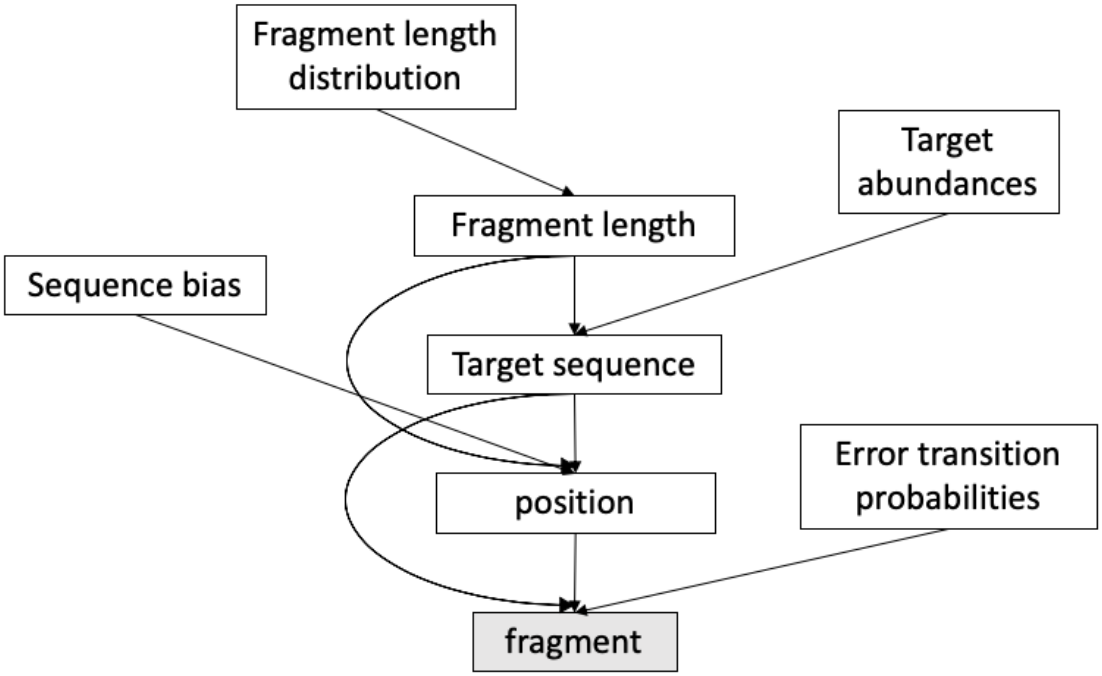


Figure 5.2: eXpress's directed graphical model. Observed variables, which can be directly observed from the reads data, are shown in grey. Latent variables, which must be inferred by the EM algorithm, are shown in white. This figure is adapted from Supplementary Figure 11 in (Roberts and Pachter, 2013).

Like RSEM, eXpress's core assumption is that its directed graphical model combined with its EM algorithm can generate accurate expression estimates. This means that assumptions about variables in the model (eg. the shape of the fragment length distribution) as well as assumptions about the relationships between variables in the model must be accurate to some degree.

### Sailfish

Sailfish is an isoform quantification tool whose performance was evaluated in my benchmark (Patro et al., 2014). Sailfish differs from the other isoform quantification tools discussed thus far in that Sailfish is an 'alignment free' tool. Instead of mapping reads, Sailfish breaks both the transcriptome and the input sequencing reads into shorter strings of a user-defined length. These shorter strings are referred to as k-mers, and as a default are of length 31.

The first step of performing isoform quantification with Sailfish is to build an index from a set of reference transcripts. Sailfish builds its index by splitting the reference transcripts into k-mers and creating a minimal perfect hash function. In simple terms, the minimal perfect hash function can be used to very quickly match a k-mer generated from reads to a location in the reference transcripts.

In the second step of quantification, Sailfish splits the input reads into k-mers and uses the minimal perfect hash function to count how many times each k-mer in its index occurs in the input reads. Sailfish then uses a conceptually similar EM algorithm to RSEM (Li and Dewey, 2011) to estimate transcript abundances based on k-mer counts.

Sailfish assumes that its k-mer hashing procedure provides a good approximation of where reads originated from in the transcriptome. The extent to which this is likely to be true depends on the k-mer length and the sequencing error rate. In addition, Sailfish assumes that its EM algorithm can generate accurate expression estimates. Sailfish also assumes that the reference transcripts passed to it represent all of the transcripts present in the transcriptome. This is potentially a problematic assumption when working with non-model organisms with poorly annotated transcriptomes.

### Salmon

Salmon is an isoform quantification tool whose performance was evaluated in my benchmark (Patro et al., 2017). Salmon has three modes - an alignment mode, a quasi mode and an SMEM mode. In the alignment mode, Salmon takes a bam

file containing aligned reads as input. The quasi and SMEM modes use a similar indexing and k-mer hashing procedure to Sailfish in place of alignment. The quasi mode is a newer mode which constructs the index and performs hashing faster than the older SMEM mode.

In addition to providing a range of alignment and alignment free modes, Salmon differs from Sailfish in that it has a more complex and sophisticated abundance estimation procedure. Salmon's abundance models account for and attempt to correct for factors such as GC-bias and positional biases in a sample-specific manner. Salmon estimates abundance in a multi-step process, beginning with a lightweight mapping step and ending with an EM algorithm to infer abundance.

Salmon assumes that its three read mapping methods provide a good approximation of where reads originated in the transcriptome. In addition, it assumes that its abundance estimation procedure can generate accurate expression estimates, meaning that it assumes that its GC bias and positional bias models accurately model these biases. When run in quasi or SMEM model, Salmon assumes that the reference transcripts passed to it represent all of the transcripts present in the transcriptome.

### Kallisto

Kallisto is an isoform quantification tool whose performance was evaluated in my benchmark (Bray et al., 2016). Kallisto uses a process described as pseudoalignment to map reads. In practice, pseudoalignment is similar to the k-mer hashing process used by Salmon and Sailfish. Like Salmon and Sailfish, Kallisto requires an index to be constructed from reference transcripts. Unlike Salmon and Sailfish, Kallisto begins indexing by constructing a data structure known as a coloured de Bruijn graph. Each node in the graph corresponds to a k-mer, and each colour corresponds to a different transcript. Nodes (k-mers) are assigned colours based on which transcripts the k-mer maps to. Contigs are linear stretches of the de Bruijn graph with the same colours and therefore the same transcripts. Once the de Bruijn graph is constructed, Kallisto generates a hash table which maps each k-mer to a contig.

During pseudoalignment, Kallisto looks up k-mers from reads in its hash table

and uses the intersect of the contigs stored in the hash table to determine which transcripts could have generated a read. Based on these pseudoalignments, Kallisto then uses an EM algorithm to infer transcript abundances.

Kallisto assumes that its pseudoalignment procedure provides a good approximation of where reads originated in the transcriptome. Kallisto also assumes that its EM algorithm can generate accurate abundance estimates. Like Salmon and Sailfish, Kallisto assumes that the reference transcripts passed to it represent all of the transcripts present in the transcriptome.

### Splatter

Splatter is a scRNA-seq counts simulator that was used with Polyester to simulate reads data (Zappia et al., 2017b; Frazee et al., 2015). Splatter originally implemented six counts simulation models, although Splatter is actively maintained and more simulation models have subsequently been added. Each simulation model has two steps. In the first step, simulation parameters are estimated from real counts data. In the second step, Splatter uses these parameters to generate a simulated counts dataset. Simulation models included in the Splatter package vary in complexity, from a very simple negative binomial model named Simple, to the more complex Splat model that accounts for mean gene expression, cell library size, the mean-variance relationship of gene expression observed in scRNA-seq and dropouts. The assumptions made by Splatter vary depending on which simulation model is used, the core assumption always being that the simulation model can generate biologically realistic counts data.

### Polyester

Polyester is a reads simulator that was used with Splatter to simulate reads data (Frazee et al., 2015; Zappia et al., 2017b). Polyester can be run using a number of different models to simulate reads. In addition, the user can specify exactly how many reads Polyester should simulate for each transcript. In my benchmark, the output of Splatter was used to dictate exactly how many reads Polyester should

simulate for each transcript. In addition, I specified that fragment length should be drawn from a normal distribution, sequencing errors should be added based on an error model that Polyester derived from a real dataset (McElroy et al., 2012) and that either no coverage bias should be simulated, or that coverage bias should be simulated based upon a cDNA fragmentation protocol. The assumptions made running Splatter with these parameters were than fragment length can be modelled as a normal distribution, the error model used was a realistic representation of error frequencies in scRNA-seq data, and that coverage bias in a SMARTer library preparation protocol could be captured with coverage bias model based upon a generic cDNA fragmentation protocol.

| Isoform Quantification Software | Input | Traditional Alignment Method? | Use of EM Algorithm? | Assumptions during isoform quantification |
|---|---|---|---|---|
| RSEM | Fastq files | Yes | Yes | • RSEM's EM model is capable of generating accurate expression estimates. |
| eXpress | Bam files | Yes | Yes | • eXpress's EM model is capable of generating accurate expression estimates. |
| Sailfish | Fastq files | No | Yes | • Sailfish's k-mer hashing procedure provides a good approximation of where reads originated in the transcriptome.<br>• Sailfish's EM model is capable of generating accurate expression estimates.<br>• The reference transcripts used to generate Sailfish's index represent all of the transcripts present in the transcriptome. |
| Salmon | Fastq files | No | Yes | • Salmon's k-mer hashing procedure provides a good approximation of where reads originated in the transcriptome (for quasi and SMEM methods)<br>• Salmon's abundance estimation models can generate accurate expression estimates.<br>• The reference transcripts used to generate Salmon's index represent all of the transcripts present in the transcriptome. |
| Kallisto | Fastq files | No | Yes | • Kallisto's pseudoalignment procedure generates a good approximation of where reads originated in the transcriptome.<br>• Kallisto's EM algorithm can generate accurate abundance estimates.<br>• The reference transcripts used to generate Kallisto's index represent all of the transcripts present in the transcriptome. |

Table 5.1: A summary table of the isoform quantification tools used in my benchmark.

### 5.1.2 Availability of data and materials

The Kolodziejczyk et al. ES cell data was accessed from the ArrayExpress database (`http://www.ebi.ac.uk/arrayexpress`) using the accession number E-MTAB-2600, as described in the Kolodziejczyk et al. paper (Kolodziejczyk et al., 2015). The BLUEPRINT data was accessed under GEO accession number GSE94676 (Adams et al., 2012). The Shekhar et al. Drop-seq data was accessed under GEO accession number GSE81905, as described in the Shekhar et al. paper (Shekhar et al., 2016).

The pipeline used to perform the BLUEPRINT benchmark, including code to reproduce figures, can be found at `https://github.com/AFS-lab/BLUEPRINT`. The pipeline used to perform the Kolodziejczyk et al. benchmark, including code to reproduce figures, can be found at `https://github.com/AFS-lab/ES_cell_pipeline`.

The pipeline to perform the Drop-seq simulation based benchmark can be found at `https://github.com/jenni-westoby/Drop-seq_pipeline`. The pipeline used to perform the systematic investigation into cell number and read depth can be found at `https://github.com/jenni-westoby/coverage_cell_number_study`. The options and parameters passed to tools used to perform simulations and isoform quantification can be found at the above links. A bug was encountered whilst using RSEM to simulate Drop-seq data. The bug was fixed and a pull request was made on the RSEM github page (`https://github.com/deweylab/RSEM/pull/79`).

### 5.1.3 Genomes

The Ensembl release 89 genome and transcriptome with 92 spike-in sequences developed by the External RNA Control Consortium (ERCC) appended were used wherever genome files in FASTQ format and/or transcriptomes in GTF format were required as input for tools in this study (Aken et al., 2017; Jiang et al., 2011). The exception to this was when isoform quantification was carried out using the BLUEPRINT and Kolodziejczyk et al. bulk RNA-seq datasets. No ERCC spike-ins were added to these datasets, so the Ensembl release 89 genome and transcriptome without spike-ins appended were used to perform this analysis. To perform simulation and isoform quantification, RSEM produces a reference which includes a

reference transcriptome in FASTQ format. This reference transcriptome produced by RSEM was used for isoform quantification tools which required a reference transcriptome in FASTQ format as input (See Github repository for code).

### 5.1.4 Data Processing Prior to Analysis

Sequencing adaptors were trimmed from the Kolodziejczyk et al. and BLUEPRINT data using Cutadapt (Martin, 2011). Reads from each cell in these datasets were aligned to the Ensembl genome release 89 using STAR (Aken et al., 2017; Dobin et al., 2013). RSeQC was used to collect alignment quality statistics for each cell (Wang et al., 2012). These statistics and the number of reads sequenced in each cell were used to remove low quality cells from each dataset (see Appendix 1, figures 7.1, 7.5 & 7.8). In addition, scater was used to plot the percentage of reads mapping to mitochondrial RNA and remove cells with greater than 10% of reads mapping to mitochondrial RNA (McCarthy et al., 2017) (See Appendix 1, figures 7.1, 7.2, 7.3, 7.4, 7.5, 7.6, 7.8, 7.9). Traditionally, Drop-seq data is not demultiplexed during gene level quantification (Macosko et al., 2015). However, with the exception of Kallisto, the tools used in this study cannot take multiplexed UMI data as input. When Kallisto does take multiplexed UMI data as input, it gives expression estimates for equivalence classes rather than for specific isoforms as output. Given that an anticipated issue with using Drop-seq for isoform quantification was that a UMI based method with 3' coverage bias may not contain enough information to resolve between different isoforms from the same gene, it was decided that the performance of Kallisto when run in this mode would not be evaluated. Instead, the Shekhar et al. dataset was demultiplexed and RSEM was used to simulate a subset of the demultiplexed cells. The performance of eXpress, Kallisto, Sailfish, Salmon and RSEM was then evaluated in the simulated cells. The cell barcodes used to demultiplex the Drop-seq data were selected by following the instructions on the Drop-seq website to generate a gene expression matrix. The barcodes were extracted from the gene expression matrix and used to demultiplex the data. Further details can be found at `https://github.com/jenni-westoby/Drop-seq_pipeline`.

### 5.1.5 Simulations

Two simulation methods were used in this study. The first method used to simulate single-cell RNA-seq data was RSEM. RSEM is an isoform quantification tool which makes use of a generative model and an expectation maximization algorithm to perform isoform quantification (Li and Dewey, 2011). When performing isoform quantification, RSEM infers values for the latent variables in its generative model in addition to estimating isoform expression. To perform simulations, RSEM takes the inferred values of the latent variable and the expression estimates and uses them in its generative model to probabilistically simulate reads. As RSEM simulates reads, it counts where in the transcriptome each of the reads came from. RSEM thus simulates reads data for which it is known how highly expressed each isoform in the transcriptome is.

For each cell in the Kolodziejczyk et al. and the BLUEPRINT datasets that passed quality control and for each of the selected cells in the Drop-seq dataset, one RSEM simulation was performed. Isoform quantification was performed on each cell and the isoform expression estimates and inferred estimates for RSEM's latent variables were used to perform the simulation. Consequently, each RSEM simulated cell used in this study was simulated using variables inferred from a real cell.

The second simulation method was based on two tools, Splatter and Polyester. Splatter is a simulation tool which takes an expression matrix of counts from a single-cell RNA-seq experiment as input and gives a simulated expression matrix of counts as output (Zappia et al., 2017a). The Splatter package in fact contains six simulation methods. To select which performed best, data was simulated using the Lun, Lun2 and Simple simulation methods. The Splat simulation method was discounted as it was unable to simulate large enough expression matrices to account for the larger number of isoforms compared with genes, and the scDD method was discounted as it simulates differential expression, and no differential expression was expected. The BASiCS method had not been implemented at the time when the simulations were performed. Based on Splatter-generated graphs, the lun2sim method, inspired by a simulation method developed by Lun & Marioni (Lun and Marioni, 2017) was

selected as it bore the closest resemblance to the real data (see Figure 2.6).

The lun2sim method was used to simulate a matrix of counts based on an expression matrix of counts from the BLUEPRINT B lymphocytes generated by Kallisto (Bray et al., 2016). The simulated expression matrix of counts was then given as input to Polyester, which simulated reads based on the lun2sim counts matrix. Simulations were performed both using Polyester's uniform coverage model and using Polyester's 3' coverage bias model. The Splatter counts matrix was converted to a matrix of TPM values, which were used as the 'ground truth' for how highly expressed each isoform was in the Polyester simulated reads data.

### 5.1.6 Post Simulation Data Processing

Reads from each cell in the datasets simulated by RSEM based on the Kolodziejczyk et al. and BLUEPRINT datasets were aligned to the Ensembl genome release 89 using STAR. RSeQC was used to collect alignment quality statistics for each cell. The alignment quality statistics and the number of reads for each simulated cell were used to remove low quality cells from each dataset (see Appendix 1, Figures 7.2, 7.3, 7.4, 7.6 & 7.9). Scater was used to plot the percentage of reads mapping to mitochondrial RNA and remove cells with greater than 10% of reads mapping to mitochondrial RNA.

### 5.1.7 Bulk RNA-seq analysis

Prior to isoform quantification, RSeQC was used to remove rRNA mapping reads from the BLUEPRINT B lymphocyte bulk RNA-seq data. The code used to generate the isoform expression matrices used in the bulk RNA-seq benchmark can be found at `https://github.com/jenni-westoby/Benchmark_Bulk_Analysis`.

### 5.1.8 Statistics

Precision and recall were used to evaluate the performance of isoform detection, whilst Spearman's Rho and the normalised root mean square error (NRMSE) were

used to evaluate the ability of tools to assign the correct magnitude of expression to each isoform. The choice of which statistics to use to evaluate performance will always be to some extent arbitrary. The precision was selected to evaluate isoform detection because the proportion of isoforms called as expressed that are truly expressed is an informative metric when determining how much confidence we can have that isoforms called as expressed are truly expressed. The recall was selected to evaluate isoform detection because the proportion of expressed isoforms called as expressed tells us how many expressed isoforms are missed by the quantification tool. Spearman's Rho was used to evaluate the ability of tools to correctly order isoform expression from the most lowly expressed to the most highly expressed, something we would hope quantification tools would be able to do well. The NRMSE was used as a measure of the error in expression estimates, which we would hope would be generally low.

The formula used to calculate the NRMSE is:

$$NRMSE = 100\frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(S_i - O_i)^2}}{sd(O)}$$

Where $N$ is the number of isoforms that could have been simulated, $S$ is the isoform expression estimates for the isoform quantification tool of interest, $O$ is the ground truth expression estimates and $sd(O)$ is the sample standard deviation of the ground truth expression estimates.

Prior to calculating the NRMSE, the ground truth and the isoform expression estimates were transformed using the formula:

$$S_{transformed} = log_2(S_{original} + 1)$$

Where $S_{original}$ was the original value of the ground truth or the expression estimate. This transformation reduces the impact of a small number of highly expressed isoforms on the value of the NRMSE.

## 5.2 Novel simulation approaches

In chapters 3 and 4, I developed two novel simulation based approaches. I present my methodology for these approaches here.

### 5.2.1 Availability of data and materials

The Kolodziejczyk et al. mESCs were accessed as described in the previous section (Kolodziejczyk et al., 2015). The hESC datasets were accessed under GEO accession number GSE85917 (Bacher et al., 2017).

My quantification pipelines, which download scRNA-seq data, perform transcript level quantification and generate an isoform-cell matrix, can be found at: `https://github.com/jenni-westoby/Isoform_Cell_Matrix_Generation`. My simulation pipeline can be found at: `https://github.com/jenni-westoby/Obstacles`.

### 5.2.2 Data processing prior to analysis

My two simulation approaches require an isoform-cell counts matrix as input. To generate isoform-cell counts matrices, I used Kallisto to quantify reads from each cell(Bray et al., 2016). For the hESC dataset, I used the Gencode human v20 transcriptome as a reference transcriptome (Frankish et al., 2019). For the mESC dataset, I used the Gencode mouse vM20 transcriptome as a reference transcriptome (Frankish et al., 2019).

### 5.2.3 Simulation Approach

My two simulation approaches are summarised as algorithms below.

```
Step 1a: Establish how many isoforms are detected for gene of interest in
  total across all cells in our scRNA-seq data.
for simulation in 1:100 do
    for i in 1:NumCells do
        for j in 1:NumExpressedIsoforms do
        end
        Step 2: Choose j isoforms to be expressed in the ith cell based on
          isoform choice model
        Step 3: Introduce dropouts based on Andrews and Hemberg's
          Michaelis-Menten model
        Step 4: Introduce isoform quantification errors
    end
    Step 5: Find mean number of isoforms per gene per cell.
end
Step 6: Plot distributions of mean number of isoforms per gene per cell (eg.
  as in Figure 3.3)
```

**Algorithm 1:** Simulation approach presented in chapter 3

```
Step 1b: Select genes for which four isoforms are detected in scRNA-seq data
for simulation in 1:100 do
    for gene in DetectedGenes do
        for j in 1:4 do
            for i in 1:NumCells do
                Step 2: Choose j isoforms to be expressed in the ith cell based
                  on isoform choice model
                Step 3: Introduce dropouts based on Andrews and Hemberg's
                  Michaelis-Menten model
                Step 4: Introduce isoform quantification errors
            end
            Step 5: Find mean number of isoforms per gene per cell.
        end
    end
end
Step 6: Plot distributions of mean number of isoforms per gene per cell (eg.
  as in Figure 4.7)
```

**Algorithm 2:** Simulation approach presented in chapter 4

Many steps are shared between the two algorithms as the two approaches are highly similar. I expand upon each step below.

**Step 1a: Establish how many isoforms are detected in total across all cells in our scRNA-seq data.**

For my simulation approach in chapter 3, I define an isoform as detected if it has more than five counts in at least two cells.

**Step 1b: Select genes for which four isoforms are detected in scRNA-seq data**

My simulation approach in chapter 4 takes an isoform-cell counts table as input. I define an isoform as detected if it has more than five counts in at least two cells. I select genes for which exactly four isoforms pass this threshold.

**Step 2: Choose $i$ isoforms to be expressed in the $j$th cell based on isoform choice model**

In this step, I probabilistically choose $i$ isoforms to be expressed in each cell, where $i$ is one, two, three or four. The default model used in chapters 3 & 4 was the Weibull model, which was used to produce all figures unless otherwise stated. In chapter 4, additional isoform choice models were also used in some figures, which are described below.

**The Weibull model**

In (Hu et al., 2017), Hu et al. found that the median frequency, $mf(k, M)$, of the $k$th dominant isoform of a gene with $M$ detected isoforms can be described as:

$$mf(k, M) = \frac{1}{k \times H_M} \exp\Big[-\Big(1 + \frac{k}{M}\Big)^2\Big]$$

where $H_M$ is the $M$th generalised harmonic number:

$$H_M = \sum_{m=1}^{M} \frac{1}{m} \exp\left[-\left(1 + \frac{m}{M}\right)^2\right]$$

In my implementation of this model of isoform choice, I first rank the isoforms in order of magnitude expression for each gene, with the most highly expressed isoform having rank 1, the second most highly expressed isoform having rank 2 and so on. I calculate magnitude of expression by summing the total number of counts across all cells for that isoform. I then use the median frequency formula above to find the predicted median frequency for each isoform. I define the probability of picking an isoform with rank $k$ for a gene with $M$ detected isoforms as:

$$p(isoform_k) = \frac{mf(k, M)}{\sum_{m=1}^{M} mf(m, M)}$$

With $M = 4$, the probabilities become [0.55, 0.28, 0.12, 0.05].

### The inferred probabilities model

In this model, I attempt to infer the probability of an isoform being chosen from its probability of being detected. The formula below relates the probability of choosing an isoform, $P(Choice)$, to its probability of being detected, $P(Detected)$:

$$\begin{aligned} P(Detection) = P(Choice) \times P(Detection|Choice) + \\ P(\neg Choice) \times P(Detection|\neg Choice) \end{aligned} \tag{5.1}$$

Where $P(\neg Choice)$ is the probability of not choosing an isoform. In practice:

$$\begin{aligned} P(Detection) = P(Choice) \times P(\neg Dropout) \times (1 - pFN) + \\ P(\neg Choice) \times pFP \end{aligned} \tag{5.2}$$

Where $P(\neg Dropout)$ is the probability that there is not a dropout, $pFN$ is the probability that there is a false negative event due to a quantification error and $pFP$

is the probability that there is a false positive event due to a quantification error. This rearranges to:

$$P(Choice) = \frac{|P(Detection) - pFP|}{|P(\neg Dropout)(1 - pFN) - pFP|}$$

In practice, I sometimes find $P(Choice)$ is greater than 1, probably because our estimation of $P(\not{D}ropout)$, $pFN$ and/or $pFP$ is inaccurate for that isoform. When this occurs, I set $P(Choice)$ equal to one. I take absolute values of the numerator and denominator to avoid negative or complex numbers, which probably also occur due to inaccurate estimation of $P(\not{D}ropout)$, $pFN$ and/or $pFP$.

In my simulations, I calculate $P(Choice)$ for each isoform from a given gene. The probability of picking a particular isoform to be expressed in our simulation is that isoform's $P(Choice)$ divided by the sum of $P(Choice)$s for that gene's isoforms.

### The cell variability model

The cell variability model is identical to the inferred probabilities model except that the probability of picking a given isoform, $i$ is allowed to vary between cells. This is acheived by sampling the probability of picking isoform $i$ in a given cell $c$, $p_ic$, from a Beta distribution, taking a similar approach to that described in (Velten et al., 2015) :

$$p_ic \sim Beta(\alpha, \beta)$$

where

$$\alpha = (\frac{1-\mu}{\sigma^2} - \frac{1}{\mu}) \times \mu^2$$

$$\beta = \alpha \times (\frac{1}{\mu} - 1),$$

where $\mu$ is the mean probability of choosing $i$ across all cells, i.e. $\mu = P(Choice)$. Based on attempts to characterise the mean-variance relationship for the probability of choosing a particular gene by Velten et al. (Velten et al., 2015), I estimate that

the sample standard deviation, $\sigma$, is approximately 0.002. I find $p_i c$ for each isoform for a given gene. In my simulations, the probability of picking isoform $i$ in cell $c$ is that isoform's $p_i c$ divided by the sum of $p_i c$s for that gene's isoforms.

### The random model

For this model, each isoform is associated a weight randomly sampled between zero and one. The probability of picking a particular isoform to be expressed in our simulation is that isoform's weight divided by the sum of all the weights for that gene's isoforms.

### The Normal model

The weights for each isoform were sampled from a truncated Normal distribution with a mean of 0.25 and a standard deviation of 0.06. This sampling was performed for each isoform in each cell. Within each cell, the probability of picking a particular isoform to be expressed in our simulation is that isoform's weight divided by the sum of all the weights for that gene's isoforms.

### The Bernoulli model

The weights for each isoform were sampled from a Bernoulli distribution with a mean of 0.25. This sampling was performed for each isoform in each cell. Within each cell, the probability of picking a particular isoform to be expressed in my simulation is that isoform's weight divided by the sum of all the weights for that gene's isoforms. If all four isoforms for a given gene had a zero weight, we set the probability of picking each isoform to 0.25.

### The p=0.25 model

The probability of choosing each isoform was always 0.25.

**Step 3: Introduce dropouts based on Andrews et al.'s Michaelis-Menten model.**

I calculate the probability of dropouts for each isoform using a Michaelis-Menten model proposed by Andrews and Hemberg (Andrews and Hemberg, 2018a). I calculate the probability of dropouts for each isoform as:

$$P(Dropout) = 1 - \frac{S}{K_M + S}$$

Where $S$ is the mean expression of that isoform across cells and $K_M$ is the Michaelis-Menten constant. To find $S$ and $K_M$ I normalise the isoform expression values by converting Counts to Counts Per Million (CPM), as suggested in the M3Drop vignette (Andrews and Hemberg, 2018a). I estimate the value of $K_M$ for each dataset by applying maximum-likelihood estimation using the equation above and the rate of dropouts and the mean expression of isoforms across the entire transcriptome.

**Step 4: Introduce quantification errors.**

Based on my benchmarking study in chapter 2 (Westoby et al., 2018b), I estimate that the probability of a false positive given an isoform has no reads mapping to it, $pFP$, is about 0.01 and the probability of a false negative given an isoform has reads mapping to it, $pFN$, is about 0.04 for Kallisto when run on full length coverage scRNA-seq data. Unless otherwise stated in the text, these were the error rates applied in my simulations. In other words, on average 1% of isoforms with no reads mapping were assigned an expressed status, and 4% of isoforms with reads mapping were assigned an unexpressed status.

**Step 5: Find mean number of isoforms per gene per cell.**

After iterating over every cell in our simulation, I sum the the number of isoforms detected in each cell and divide by the number of cells to find the mean number of detected isoforms per gene per cell.

| |
|---|
| Klf4 |
| Pou5f1 |
| Tbx3 |
| Jarid2 |
| Myc |
| Stat3 |
| Tcf3 |
| Esrrb |
| Nanog |
| Nr0b1 |
| Sall4 |
| Sox2 |
| Zfp42 |
| Zfp281 |
| Zfx |

Table 5.2: Pluripotency factor genes.

**Step 6: Plot distributions of mean number of isoforms per gene per cell**

To make my simulation results figures shown in chapter 3, I plot the distributions of the mean number of detected isoforms for each number of expressed isoforms, from 1 to the total number of isoforms detected. A vertical black line representing the mean number of isoforms detected per cell is drawn on the same plot (for example, see Figure 3.3).

The distributions plotted in chapter 4 differ. In chapter 3, the distributions generated are for a single gene of interest. In chapter 4, the distributions produced are for every gene in which 4 isoforms were detected in the real scRNA-seq data (eg. see Figure 4.7).

### 5.2.4 Genes investigated in chapter 3

I ran my model on 15 pluripotency factor genes. The 15 genes are shown in Table 5.2.

### 5.2.5 Mixture Modelling

In my mixture modelling experiments in chapter 4, I begin by fitting log normal distributions to each of our simulation distributions and to the distribution of mean isoforms detected for genes with four detected isoforms in the real data. I then use expectation maximisation to estimate the mixing fraction of each of the simulated distributions in the real distribution. In my expectation step, I calculate the probability that each data point belongs to a given distribution, which I refer to as the responsibility. The responsibility for the $i$th mean number of isoforms per gene per cell and the $c$th simulation distribution is:

$$r_i c = \frac{k_c \times LN(x_i|\mu_c, \sigma_c)}{\sum_{j=1}^{j=4} k_j \times LN(x_i|\mu_j, \sigma_j)}$$

where $k$ is the mixing fraction, $x_i$ is the $i$th mean number of isoforms per gene per cell and $LN(x_i|\mu_c, \sigma_c)$ is the probability density function for the log normal with mean $\mu_c$ and variance $\sigma_c^2$. The maximisation function for the mixing fraction is:

$$k_c = \frac{\sum_i r_i c}{n}$$

Where $n$ is the number of datapoints in $r_i c$. Note that I only perform expectation maximisation for the mixing fractions of the distributions and not for the means or standard deviations.

## Overlap Fraction

The overlap fraction is the proportion of isoforms detected in our simulations that were expressed in the ground truth. The formula for the overlap fraction is:

$$OverlapFraction = \frac{|GroundTruth \cap Detected|}{|GroundTruth|}$$

Where $GroundTruth$ is the set of isoforms that are expressed in the ground truth, and $Detected$ is the set of isoforms that are detected in our simulations. The overlap

fractions reported in all Figures & Supplementary Figures are the mean overlap fractions for each gene, averaged across all of the simulated cells in that simulation round.

## Downsampling

Random downsampling of reads in chapter 4 was performed using seqtk (Li, 2013).

# 6

# Discussion

> *I once wrote a lecture for Manchester University called 'Moments of Discovery' in which I said that there are two moments that are important. There's the moment when you know you can find out the answer and that's the period you are sleepless before you know what it is. When you've got it and know what it is, then you can rest easy.*
>
> – Dorothy Hodgkin (Hodgkins)

The main goal of my thesis was to assess the extent to which it is feasible to study alternative splicing using scRNA-seq. I have established that from a software perspective, existing isoform quantification tools perform well when run on scRNA-seq data. However, I have also established that this alone is not sufficient to enable splicing to be accurately studied at a cellular resolution. Without a better understanding of dropouts and isoform choice, our ability to accurately detect isoforms in individual cells is poor. In the final chapter of my thesis, I will discuss the main findings from my PhD and consider how the field could move forward.

## 6.1 My benchmarking study demonstrated that isoform quantification tools designed for bulk RNA-seq perform well when run on scRNA-seq

Most scRNA-seq studies quantify reads at the level of genes rather than isoforms (Stegle et al., 2015), partly due to uncertainty over whether appropriate tools to quantify reads at the isoform level exist for scRNA-seq. My benchmark demonstrated that Kallisto, Salmon, Sailfish and RSEM perform almost as well when run on scRNA-seq as when run on bulk RNA-seq. Thus, it seems unlikely that a lack of appropriate isoform quantification software could prevent us from studying alternative splicing using scRNA-seq.

Important insight came when data was simulated based on Drop-seq rather than SMARTer or SMART-seq2 data. The performance of almost all isoform quantification tools was so poor when run on simulated Drop-seq data that accurate quantification or detection of isoforms was impossible. This indicates that choice of library preparation protocol matters if the goal of an scRNA-seq experiment is to study alternative splicing. I considered three library preparation protocols in my benchmark. Further work evaluating the suitability of other popular library preparation protocols for studying splicing is likely to be valuable to the single cell community.

An unexpected finding from my benchmark was that the precision of isoform detection peaks at around 1-2 million reads per cell. I hypothesise that this occurs because RSEM does not substantially increase the number of isoforms it expresses per cell beyond 1-2 million reads per cell. As the number of reads simulated increases beyond 2 million reads per cell, the opportunity to increase the number of true positives (expressed isoforms which are called as expressed) is therefore limited. However, if there is a fixed probability that a randomly selected read will be misassigned to the wrong isoform, as the number of reads increases we would expect the number of false positives to continue to rise. As the formula for the precision is:

$$Precision = \frac{No.TruePositives}{No.TruePositives + No.FalsePositives}$$

we would therefore expect the precision to decrease at read depths greater than 1-2 million reads per cell.

As previously stated, I believe that the position of the peak in precision at 1-2 million reads per cell is likely a simulation artefact of RSEM. However, it is a fact that cells express a finite number of isoforms. Therefore if my hypothesis is correct, I would predict that there is also a peak in the precision of isoform detection in real scRNA-seq data. Indeed, if I am correct, I predict that this phenomenon is not limited to scRNA-seq or to isoform level quantification, but that this phenomenon would also occur for other sequencing technologies (eg. bulk RNA-seq) and for gene level quantification. I predict that the location of the precision peak would be determined by the number of isoforms (or genes, in the case of gene level quantification) expressed per cell. It has been established that for gene level quantification, the number of genes detected plateaus at 1 million reads per cell (Wu et al., 2014; Ziegenhain et al., 2017). Therefore my prediction is that the peak in the precision of gene detection occurs at around 1 million reads per cell. Establishing the read depth at which isoform detection plateaus could give insight into the read depth at which the precision of isoform detection is likely to peak.

I make a lot of predictions in the previous paragraph, so an obvious question is how these predictions could be tested. One way to test my predictions would be by performing a scRNA-seq experiment in which spike-in mixtures, composed of a mixture of a known number transcript species spiked at known concentrations, are sequenced in place of cells. By sequencing the spike-in mixture at varying depths, it could be established whether the number of transcript species predicts the position of the peak in isoform detection precision. Datasets such as these could be used to build a model to predict the precision of isoform detection at various read depths. This could give insight into what statistical frameworks could be used to best correct for an increasing proportion of false positives as read depth increases.

My observation that the precision of isoform detection peaks raises a lot of unresolved issues. The experiments that gave this insight were only performed using Salmon. It is possible that other isoform quantification tools may correct for the increased proportion of false positives at higher read depths more effectively. Establishing whether the precision peaking phenomenon is common to all isoform quantification tools, or if it is a quirk of Salmon, would be of relevance to the bioinformatics community. If it is a common phenomenon, new statistical frameworks should be generated to correct for it, potentially based on results from spike-in sequencing data as described above. Correcting for the peak in precision is especially relevant when analysing data which has been sequenced to very high depths. An important implication of the precision peaking phenomenon is that many of the 'rare' or 'lowly expressed' genes only detected when sequencing at very high depths and/or with vast numbers of cells are likely to be false positives.

Although I believe that my benchmarking study has generated many useful insights, no simulation based approach is without limitations. I have carefully evaluated the similarities and differences between the simulated and real data, but the possibility remains that the simulated data in some way fails to capture some meaningful aspect of the real data. However, given that the results of my benchmark remain consistent across two unrelated simulation methods, I am optimistic that the results of my benchmark are sufficiently reliable to be applied to real data.

An important aspect of my simulation based approach is that the 'ground truth' is based on where reads originated from in the transcriptome. In other words, my benchmark evaluated the ability of isoform quantification tools to correctly determine where reads originated from in the transcriptome. This is an important ability for an isoform quantification tool to have, however it should be recognised that scRNA-seq has a high frequency of technical dropouts and PCR amplification bias. Because my benchmark only considers the ability of quantification software to quantify the reads that are present, I entirely ignore the issues arising from all the reads from expressed isoforms that are absent due to dropouts and the issues arising from amplification bias. The only way to benchmark the ability of isoform quantification tools to correctly infer the original expression of isoforms in cells would be if an scRNA-seq

experiment were carried out where the exact expression of isoforms was known at the moment that the RNA was extracted from the cells. At present, no technology exists that could generate this type of data. Consequently, I made identifying isoform quantification tools that can accurately quantify the reads that are present the goal of my benchmarking study, and consider how best to correct for technical noise such as dropouts to be a separate issue.

## 6.2 Initial attempts to determine how many isoforms are produced per gene per cell gave uninterpretable results

How many isoforms does a gene produce in a cell? This deceptively simple question holds a fundamental place in molecular biology. I began my attempts to investigate how many isoforms a gene produces in a cell by considering genes for which two isoforms are detected in bulk RNA-seq, and asking how many isoforms are detected per cell using scRNA-seq. I found that I frequently failed to detected gene expression at all, and when I did detect gene expression, it was more common to only detect the expression of only one isoform. This is consistent with two hypotheses. The first hypothesis is that gene expression genuinely is heterogeneous between cells, and that it is rare for cells to simultaneously express two isoforms from the same parent gene. The second hypothesis is that we commonly fail to detect gene and isoform expression due to technical dropouts. Without a better understanding of how best to model dropouts, it is challenging to determine to what extent each of these hypotheses are true. Similarly, my later observation that two isoforms are more frequently detected for more highly expressed genes is consistent with the hypothesis that more highly expressed genes undergo more splicing, and is also consistent with the hypothesis that dropouts are more prevalent for less expressed transcripts. Again, it is not currently possible to determine which of these hypotheses is most accurate without a more sophisticated understanding of dropouts. Therefore, I conclude that at present, the results of the experiment are uninterpretable.

As dropouts made the results of my previous experiment uninterpretable, the approach taken in my next experiment explicitly accounted for dropouts using a Michaelis-Menten model developed by Andrews and Hemberg (Andrews and Hemberg, 2018a). However, the model I developed to predict how many isoforms are produced per gene per cell failed at the first hurdle by predicting that more isoforms were produced per cell than could be detected across all cells. This prediction is possible if the rate of dropouts is extremely high, but given that my model's predictions are not supported by matched bulk RNA-seq data either, this explanation seems unlikely. One solution would be to test the predictions using smFISH. Whilst this would establish whether the model makes accurate predictions or not, it would be expensive, technically challenging and time consuming to carry out smFISH experiments for a large enough number of genes to confidently evaluate the model's accuracy. Ultimately I think smFISH experiments resolving the number of isoforms produced from a gene in individual cells would be hugely valuable from both a molecular biology and single cell perspective, however they go beyond the scope of this thesis. A more realistic solution could be to modify my model so that it makes more 'realistic' predictions. Modifying my model is clearly feasible, however whether it would be useful is another question. Altering a model with the goal of making its predictions more 'realistic' without knowing how many isoforms are produced per gene per cell in reality is likely to create a model that reflects my own biases more than anything else.

The main outcome from these two experiments was the realisation that a lack of understanding about technical noise in scRNA-seq can make a biological interpretation of scRNA-seq data challenging, if not impossible. If more was understood about dropouts, and how to correct for dropouts in individual cells, the two experiments discussed above would become far easier to interpret. To address this lack of understanding about the extent to which technical noise confounds scRNA-seq experiments, I embarked on a final simulation based study. In my final study, I investigated the extent to which certain variables confound our ability to detect isoforms in individual cells.

## 6.3 Dropouts are a major obstacle to studying alternative splicing using scRNA-seq

In the final Results chapter of my thesis, I developed a novel simulation based approach and considered to what extent certain variables confound our ability to study alternative splicing in individual cells. I considered one biological and two technical variables. The technical variables considered were dropouts, which I found to be a major confounder, and isoform quantification errors. Isoform quantification errors could theoretically be entirely or almost entirely removed as a source of technical noise in scRNA-seq experiments, especially if long read technologies improve. Therefore, establishing whether quantification errors are a major confounder is of interest to the field because there are clear actions that could be taken to reduce or remove confounding. In practice however, I found that isoform quantification errors were a relatively minor confounder. Indeed, even when all quantification errors were removed from my simulations, substantial confounding remained. This illustrates that improvements in the accuracy of isoform quantification would not be sufficient to enable accurate splicing analyses using scRNA-seq.

In addition to dropouts and quantification errors, I asked whether different models of isoform choice, in which each cell 'chooses' which isoforms should be expressed, have an impact on my simulation results. I found that different models of isoforms choice significantly altered my simulation results, indicating that there could be value in including it in future models for studying alternative splicing using scRNA-seq. Insight into the cellular process of isoform choice could be gained from smFISH experiments, which would enable more accurate models of isoform choice to be built. My results suggest that we are less able to detect isoforms in individual cells when cells use models of isoform choice in which isoforms with a high probability of dropout are frequently chosen. Determining whether lowly expressed isoforms with high dropout probabilities are frequently expressed by cells will therefore be important and give an indication of the extent to which we are currently able to detect isoforms in individual cells using scRNA-seq.

I considered three potentially confounding variables in my simulations, however

I recognise that other sources of biological and technical noise exist in scRNA-seq that could potentially confound splicing experiments. Because I focus on detecting rather than quantifying isoforms, I do not consider PCR amplification bias, which in theory should not directly impact on isoform detection. With good experimental design batch effects can often be avoided in scRNA-seq experiments, so I do not consider batch effects in my simulations either. Biological dropouts, for example due to transcriptional bursting or cell type specific expression of isoforms, are a potential confounder when studying splicing using scRNA-seq. Distinguishing between biological and technical dropouts is challenging and little work has been done to resolve between biological and technical dropouts in individual cells in scRNA-seq data. Andrews and Hemberg's Michaelis-Menten model was originally designed to identify genes with more dropouts than expected based on the gene's mean expression, the rationale being that the excess dropouts must be biological dropouts (Andrews and Hemberg, 2018a). However, to build an evidence based model of biological dropouts for use in my simulations or related approaches, this work would need to be substantially extended. Additionally, models have previously been built which attempt to capture transcriptional bursting behaviour from scRNA-seq data (Kim and Marioni, 2013; Kim et al., 2015; Larsson et al., 2019; Ochiai et al., 2019). However these models do not account for cell type specific expression of isoforms, thus are not comprehensive models of all biological dropouts. Additionally, many of these models were not validated using any orthogonal method (eg. smFISH) to test their predictions, meaning their biological relevance is unclear. Therefore, in my simulations I do not consider biological dropouts due to a lack of appropriate models of biological dropouts. I am hopeful that further work might be done in this space, as I think it would be of great value to the community.

In the final Results chapter of my thesis, I established that alternative splicing analyses using scRNA-seq are currently confounded by dropouts and a lack of understanding about the cellular process of isoform choice. As a field, we do not currently know how best to correct for these confounders. Consequently, I conclude that at present I cannot recommend attempting alternative splicing using scRNA-seq data. So how should we move forwards?

## 6.4 Future Directions

A recurring theme throughout my thesis has been that dropouts can be a substantial confounder when attempting to detect isoforms in individual cells, but it is unknown how best to correct for this confounder. There are three very different approaches to attempt to solve this problem. The first is to build models of technical and biological dropouts that would enable dropouts to be resolved in individual cells in a high confidence manner. This is essentially isoform-level imputation. Multiple tools have been developed that attempt to impute scRNA-seq data quantified at the gene level (Wagner et al., 2017; Li and Li, 2018; Huang et al., 2018; Gong et al., 2018; van Dijk et al., 2018; Eraslan et al., 2019), however their poor performance in a recent benchmark illustrates that this is a highly challenging problem that we are some way from solving (Andrews and Hemberg, 2018b). In general, bioinformatics dropout based methods such as imputation and Andrews and Hemberg's Michaelis-Menten model have focused on correcting for dropouts for applications such as clustering and feature selection. Different approaches might be required when attempting to resolve dropouts in individual cells - for example, factors such as the cell's library size and physical size might need to be accounted for. Attempting to develop such approaches could enable more accurate splicing analyses to be performed in future. However, this is an extremely challenging problem to solve, as illustrated by the poor performance of existing imputation tools (Andrews and Hemberg, 2018a).

The second approach that could be taken to attempt to solve confounding effects caused by dropouts would be to increase the capture efficiency of scRNA-seq. It is hypothesised that dropouts occur due inefficiencies in the enzymatic process of reverse transcription (Kharchenko et al., 2014). If this hypothesis is correct, improvements to the efficiencies of the enzymatic reactions that occur during library preparation could reduce the frequency of dropouts in scRNA-seq. Whilst this thesis was being written, a new library preparation protocol called SMART-seq3 was released (Hagemann-Jensen et al., 2019). One of the stated improvements in SMART-seq3 relative to SMART-seq2 was that the efficiency of several enzymatic reactions in library preparation had been improved, which could theoretically improve the cap-

ture efficiency of SMART-seq3. Indeed, in their preprint, Hagemann-Jensen et al. showed that SMART-seq3 detected more genes per cell on average than SMART-seq2, which would be consistent with an increased capture efficiency. Hagemann-Jensen et al. also claimed that SMART-seq3 on average detected an estimated 69% of the molecules detected from four moderately expressed genes using smFISH. However, only four genes were reported and it is unclear how these estimates were generated, so these claims should perhaps be taken with a pinch of salt. An independent benchmark of library preparation protocols, including SMART-seq3, is required to determine whether the capture efficiency of SMART-seq3 is genuinely elevated relative to other library preparation methods. If SMART-seq3 truly does have a higher capture efficiency, it could play an important role in enabling the detection of isoforms in individual cells.

The final approach to solving confounding factors caused by dropouts in scRNA-seq data is to use a different technology to detect isoforms in individual cells. smFISH is the most obvious candidate. Whilst smFISH has traditionally been a low throughput technology which struggles to resolve between isoforms, recent and future improvements in throughput (Eng et al., 2019; Moffitt et al., 2016) and techniques to resolve between similar molecules (Levesque et al., 2013) could make a high throughput study of how many isoforms are expressed per gene per cell increasingly feasible. An smFISH dataset resolving the number of isoforms detected per gene per cell for a hundred or so genes would be hugely valuable to the scRNA-seq community. This dataset could be used as a ground truth dataset to benchmark scRNA-seq methods for inferring isoform number. Additionally, such an smFISH dataset could be used to train and test machine learning approaches, which are currently impossible due to a lack of training data.

An important point to recognise is that my simulation based approach only focussed on isoform detection. Establishing the relative magnitude of expression of isoforms is likely to be of interest to many researchers, however simply detecting isoforms accurately is currently problematic. Therefore, accurately inferring the relative magnitude of expression of isoforms in individual cells is not yet feasible in my view. Furthermore, the two library preparation protocols which performed well in my

benchmarking study (SMARTer and SMART-seq2) do not add UMIs to transcripts and so suffer from PCR amplification bias. This is likely to substantially confound attempts to infer magnitude of expression. SMART-seq3 does add UMIs to some reads, however Hagemann-Jensen et al. demonstrate that the UMI containing reads have substantial bias towards the 5' end of the transcript (Hagemann-Jensen et al., 2019). This 5' bias is likely to make the detection and quantification of isoforms that differ at their 3' end challenging.

scRNA-seq is a very dynamic field, and many researchers are actively working on new technological developments. In the next section, I consider whether up and coming developments in scRNA-seq technologies could improve the feasibility of studying splicing in the future.

## 6.5 scRNA-seq technologies on the horizon

As the timeline in Figure 1.5 illustrates, many scRNA-seq technologies have been developed in the past decade. However, further developments are on the horizon. In this section, I will discuss several exciting technological developments that are on the verge of becoming practical and consider whether in the future, they could play a role in enabling splicing to be studied using scRNA-seq.

### 6.5.1 SMART-seq3

A preprint for a new scRNA-seq library preparation protocol called SMART-seq3 was recently released (Hagemann-Jensen et al., 2019). SMART-seq3 has been widely described as the first technology to combine full length reads and reads containing UMIs. However, it is important to recognise that full length reads containing UMIs are not generated by this protocol. Rather, this protocol generates a set of UMI containing reads which are heavily biased towards the 5' end of the transcript, and a set of full length reads that do not contain UMIs and have relatively a uniform coverage.

SMART-seq3 is certainly a very novel approach, however it is not immediately

obvious what its applications will be. In their preprint, Hagemann-Jensen et al. propose that the UMI containing reads could be used to infer isoform structure and allelic information. However, as the UMI containing reads are 5' biased, it is unclear how effective this approach would be, especially compared to potential developments in long read technologies. Given the mix of UMI and non-UMI containing reads generated by this protocol, research into the best bioinformatics methods for analysing this data may be required to determine how best to use data generated by the new protocol. In addition, independent benchmarks comparing SMART-seq3 to other library preparation protocols would be valuable. In their preprint, Hagemann-Jensen et al. attempt to improve the efficiency of enzymatic reactions in their library preparation protocol. If the enzyme reaction efficiency has been improved, the capture efficiency of SMART-seq3 could be higher than that of other library preparation protocols. The capture efficiency of SMART-seq3 should be independently investigated and compared to other scRNA-seq technologies to establish whether it truly is elevated. If SMART-seq3 does have a higher capture efficiency than other scRNA-seq technologies, this could have important implications for the feasibility of studying splicing using scRNA-seq. I have established in this thesis that dropouts are a major confounder when trying to study splicing with scRNA-seq. A reduction in the rate of technical dropouts could meaningfully increase the feasibility of accurately detecting the number of isoforms in individual cells.

### 6.5.2 Long read scRNA-seq

At the time of writing, most sequencing experiments use Illumina sequencing. However, other sequencing platforms are available. A distinct advantage of some sequencing platforms is that they enable longer reads to be sequenced than is currently possible using Illumina. If we are interested in studying alternative splicing this is exciting news. Longer reads would cover more of the length of each transcript, theoretically enabling more accurate transcript identification. If the entire length of the transcript could be covered by a single read, in theory that transcript could be identified with perfect accuracy.

In practice, existing long read sequencing platforms such as PacBio and Oxford Nanopore have a much higher base calling error rate than Illumina (Koren et al., 2012; Rang et al., 2018; Fu et al., 2019). Thus, it is unclear whether the accuracy of transcript identification would overall be higher or lower using this platform compared with using Illumina. Independent comparative benchmarks are needed to address this. In addition, improvements in base calling accuracy are likely to improve the accuracy of transcript identification. However, I note that in my simulations in chapter 4, when the isoform quantification error rate was reduced to zero, our ability to detect isoforms in individual cells remained poor. This is most likely due to confounding caused by dropouts. Long read technologies would represent an important advance in scRNA-seq and could dramatically improve transcript identification. But unless long read technologies are accompanied with improvements in the capture efficiency of scRNA-seq or better methodologies to correct for dropouts, long read technologies will not solve all of the issues currently faced when trying to study splicing with scRNA-seq.

To date, a small number of long read scRNA-seq publications exist (Singh et al., 2019; Gupta et al., 2018; Karlsson and Linnarsson, 2017; Lebrigand et al., 2019). However, long read technologies are still young. Consequently, challenges relating to cost, technical difficulties, a lack of well established protocols and uncertainty over data quality mean that few labs have attempted long read scRNA-seq to date. In addition, at present the read throughput of long read platforms is too low to enable meaningful isoform detection and quantification across a large number of cells (Arzalluz-Luque and Conesa, 2018). Notwithstanding, long read scRNA-seq is of interest to the single cell community and can be expected to develop and grow in years to come.

### 6.5.3 Spatial transcriptomics

Spatial transcriptomics is a field in which information about where cells are located in tissues is linked with information about what genes or isoforms they transcribe. Given that one of the most common applications of scRNA-seq is cell identification,

there is interest in the single cell community in incorporating spatial information with scRNA-seq analyses, as spatial information could be used to validate predicted cell identities and to provide new insight into how tissues are organised at the cellular level. From a splicing perspective, spatial transcriptomics could enable researchers to correlate differential splicing with a cell's physical location. This would represent a genuine advance on what is currently possible using scRNA-seq.

Broadly speaking, spatial transcriptomics falls into two main categories - FISH based spatial transcriptomics and scRNA-seq based spatial transcriptomics. From a validation perspective, it is valuable that both FISH and scRNA-seq based approaches exist. FISH based approaches could provide an orthogonal means of validating splicing and other biological predictions from scRNA-seq based approaches. Traditionally FISH based spatial transcriptomics approaches were relatively low throughput in terms of the number of genes that could be assayed. However throughput has been improving. A recently developed FISH based spatial transcriptomic technology, seqFISH+, was used to profile 10,000 genes in nearly 3,000 cells (Eng et al., 2019).

In recent years, a series of scRNA-seq based spatial transcriptomic methods have been developed. The details of the protocols vary, however the general principle is that a thin tissue section is prepared, then is laid over a slide coated with beads or probes which capture RNA (Rodriques et al., 2019; Vickovic et al., 2019; Ståhl et al., 2016). These procedures are not truly single cell as the beads or probes often capture RNA from more than one cell. Nonetheless, they represent a genuine breakthrough in our ability to link spatial and transcriptomic information. Improvements to this technology, such as only capturing transcriptomic information from one cell, reduced cost and simpler protocols, could lead to spatial transcriptomics becoming increasingly widespread in the future.

## 6.6 Alternative splicing and scRNA-seq: conclusions from my feasibility assessment

I began my feasibility assessment by asking whether existing isoform quantification tools give accurate results when run on scRNA-seq data. I found that in general, existing isoform quantification tools perform well when run on scRNA-seq data, provided that the data has full length reads and the sequencing depth is moderately high for each cell. However, when I attempted to address biological questions using tools that performed well in my benchmark, I found that a lack of knowledge about what confounders were present in the scRNA-seq data made my results impossible to interpret.

I therefore continued my feasibility assessment by investigating the extent to which various technical and biological factors confounded splicing analyses using scRNA-seq. I found that dropouts are a major technical confounder when attempting to detect isoforms in individual cells. Isoform quantification errors were a much lesser confounder. Importantly, even when isoform quantification is error free, substantial confounding remains. This indicates that perfect isoform quantification alone is insufficient to enable accurate splicing analyses. My results indicate that isoform choice can impact on our ability to detect isoforms in individual cells, so should be accounted for in future splicing analyses.

At present, it is unclear whether the capture efficiency of scRNA-seq could be increased and accurate methodologies to correct for technical dropouts do not exist. In addition, little is known about the cellular mechanisms of isoform choice for most genes. As I have shown that dropouts and a lack of knowledge about isoform choice confound our ability to analyse splicing using scRNA-seq, I conclude that at present, it is often not feasible to accurately analyse alternative splicing using scRNA-seq. However, I am optimistic that with a combination of bioinformatics methods development, wet lab experiments and improvements to scRNA-seq technologies, it may become feasible to study alternative splicing using scRNA-seq in the future.

## 6.7 A methods-driven approach to biology

As a bioinformatician, I have often been encouraged not to lose sight of the biology. However, as I have become more experienced over the course of my PhD, I have come to question how much biology it is currently possible to see through the lens of genomics. Bioinformatics is not, and should not be regarded as, a magic wand that removes all of the flaws and noise in genomics data. Bioinformatics software can be blindly applied to genomics data and it will often produce an answer. We have many roles as scientists, and one of those roles should be to question whether it was appropriate to apply that bioinformatics methodology to that dataset. If it was not, the answer that was generated should be questioned.

The methods-driven approach I took in chapter 4 of my thesis was in some respects entirely detached from the underlying biological process of isoform choice in individual cells. The four situations that I simulated, in which each cell expressed exactly one, two, three or four isoforms, most likely bear no resemblance to the cellular regulation of isoform expression in reality. Nonetheless, I regard chapter 4 as the most important chapter of my thesis. Chapter 4 delivered insight into what biological questions currently can and can not be answered using scRNA-seq. It is my view that genomics data should only be used to answer biological questions if there is a reasonable expectation that the answer will be accurate. As I have demonstrated in this thesis, determining whether an accurate answer is obtainable can require considerable research. I am hopeful that similar approaches to the approach taken in chapter 4 will be used more frequently in future.

Academia can be slow to change and I am a very minor player. However, if I could persuade a few scientists to question and systematically evaluate the bioinformatics methodologies and approaches used in their research, I would feel that the last three years were exceptionally well spent.

# Bibliography

D. Adams, L. Altucci, S. E. Antonarakis, J. Ballesteros, S. Beck, A. Bird, C. Bock, B. Boehm, E. Campo, A. Caricasole, F. Dahl, E. T. Dermitzakis, T. Enver, M. Esteller, X. Estivill, A. Ferguson-Smith, J. Fitzgibbon, P. Flicek, C. Giehl, T. Graf, F. Grosveld, R. Guigo, I. Gut, K. Helin, J. Jarvius, R. Küppers, H. Lehrach, T. Lengauer, r. Lernmark, D. Leslie, M. Loeffler, E. Macintyre, A. Mai, J. H. A. Martens, S. Minucci, W. H. Ouwehand, P. G. Pelicci, H. Pendeville, B. Porse, V. Rakyan, W. Reik, M. Schrappe, D. Schübeler, M. Seifert, R. Siebert, D. Simmons, N. Soranzo, S. Spicuglia, M. Stratton, H. G. Stunnenberg, A. Tanay, D. Torrents, A. Valencia, E. Vellenga, M. Vingron, J. Walter, and S. Willcocks. BLUEPRINT to decode the epigenetic signature written in blood. *Nature Biotechnology*, 30(3):224–226, mar 2012. ISSN 1087-0156. doi: 10.1038/nbt.2153. URL `http://www.nature.com/doifinder/10.1038/nbt.2153`.

D. E. Agafonov, B. Kastner, O. Dybkov, R. V. Hofele, W.-T. Liu, H. Urlaub, R. Lührmann, and H. Stark. Molecular architecture of the human u4/u6.u5 tri-snRNP. *Science*, 351(6280):1416–1420, mar 2016. doi: 10.1126/science.aad2085. URL `http://dx.doi.org/10.1126/science.aad2085`.

B. L. Aken, P. Achuthan, W. Akanni, M. R. Amode, F. Bernsdorff, J. Bhai, K. Billis, D. Carvalho-Silva, C. Cummins, P. Clapham, L. Gil, C. G. Girón, L. Gordon, T. Hourlier, S. E. Hunt, S. H. Janacek, T. Juettemann, S. Keenan, M. R. Laird, I. Lavidas, T. Maurel, W. McLaren, B. Moore, D. N. Murphy, R. Nag, V. Newman, M. Nuhn, C. K. Ong, A. Parker, M. Patricio, H. S. Riat, D. Sheppard, H. Sparrow, K. Taylor, A. Thormann, A. Vullo, B. Walts, S. P. Wilder, A. Zadissa,

M. Kostadima, F. J. Martin, M. Muffato, E. Perry, M. Ruffier, D. M. Staines, S. J. Trevanion, F. Cunningham, A. Yates, D. R. Zerbino, and P. Flicek. Ensembl 2017. *Nucleic Acids Research*, 45(D1):D635–D642, jan 2017. doi: 10.1093/nar/gkw1104. URL http://dx.doi.org/10.1093/nar/gkw1104.

S. Andrews. Fastqc. Jun 2015. URL https://qubeshub.org/resources/fastqc.

T. S. Andrews and M. Hemberg. M3Drop: Dropout-based feature selection for scRNASeq. *Bioinformatics*, dec 2018a. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty1044. URL http://fdslive.oup.com/www.oup.com/pdf/production_in_progress.pdf.

T. S. Andrews and M. Hemberg. False signals induced by single-cell imputation. *F1000Research*, 7:1740, nov 2018b. doi: 10.12688/f1000research.16613.2. URL http://dx.doi.org/10.12688/f1000research.16613.2.

M. Anokhina, S. Bessonov, Z. Miao, E. Westhof, K. Hartmuth, and R. Lührmann. RNA structure analysis of human spliceosomes reveals a compact 3D arrangement of snRNAs at the catalytic core. *The EMBO Journal*, 32(21):2804–2818, oct 2013. doi: 10.1038/emboj.2013.198. URL http://dx.doi.org/10.1038/emboj.2013.198.

Á. Arzalluz-Luque and A. Conesa. Single-cell RNAseq for the study of isoforms-how is that possible? *Genome Biology*, 19(1):110, aug 2018. doi: 10.1186/s13059-018-1496-z. URL http://dx.doi.org/10.1186/s13059-018-1496-z.

R. Bacher and C. Kendziorski. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biology*, 17:63, apr 2016. doi: 10.1186/s13059-016-0927-y. URL http://dx.doi.org/10.1186/s13059-016-0927-y.

R. Bacher, L.-F. Chu, N. Leng, A. P. Gasch, J. A. Thomson, R. M. Stewart, M. Newton, and C. Kendziorski. SCnorm: robust normalization of single-cell RNA-seq data. *Nature Methods*, 14(6):584–586, jun 2017. doi: 10.1038/nmeth.4263. URL http://dx.doi.org/10.1038/nmeth.4263.

N. Behzadnia, M. M. Golas, K. Hartmuth, B. Sander, B. Kastner, J. Deckert, P. Dube, C. L. Will, H. Urlaub, H. Stark, and R. Lührmann. Composition and three-dimensional EM structure of double affinity-purified, human prespliceosomal a complexes. *The EMBO Journal*, 26(6):1737–1748, mar 2007. doi: 10.1038/sj.emboj.7601631. URL http://dx.doi.org/10.1038/sj.emboj.7601631.

S. M. Berget and B. L. Robberson. U1, u2, and u4/u6 small nuclear ribonucleoproteins are required for in vitro splicing but not polyadenylation. *Cell*, 46(5):691–696, aug 1986. doi: 10.1016/0092-8674(86)90344-2. URL http://dx.doi.org/10.1016/0092-8674(86)90344-2.

S. M. Berget, C. Moore, and P. A. Sharp. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(8):3171–3175, aug 1977. doi: 10.1073/pnas.74.8.3171. URL http://dx.doi.org/10.1073/pnas.74.8.3171.

K. Bertram, D. E. Agafonov, W.-T. Liu, O. Dybkov, C. L. Will, K. Hartmuth, H. Urlaub, B. Kastner, H. Stark, and R. Lührmann. Cryo-EM structure of a human spliceosome activated for step 2 of splicing. *Nature*, 542(7641):318–323, feb 2017. doi: 10.1038/nature21079. URL http://dx.doi.org/10.1038/nature21079.

S. Bessonov, M. Anokhina, A. Krasauskas, M. M. Golas, B. Sander, C. L. Will, H. Urlaub, H. Stark, and R. Lührmann. Characterization of purified human bact spliceosomal complexes reveals compositional and morphological changes during spliceosome activation and first step catalysis. *RNA (New York)*, 16(12):2384–2403, dec 2010. doi: 10.1261/rna.2456210. URL http://dx.doi.org/10.1261/rna.2456210.

A. Bindereif and M. R. Green. An ordered pathway of snRNP binding during mammalian pre-mRNA splicing complex assembly. *The EMBO Journal*, 6(8):2415–2424, aug 1987. doi: 10.1002/j.1460-2075.1987.tb02520.x. URL http://dx.doi.org/10.1002/j.1460-2075.1987.tb02520.x.

D. L. Black, B. Chabot, and J. A. Steitz. U2 as well as u1 small nuclear ribonucleoproteins are involved in premessenger RNA splicing. *Cell*, 42(3):737–750, oct 1985. doi: 10.1016/0092-8674(85)90270-3. URL http://dx.doi.org/10.1016/0092-8674(85)90270-3.

M. Blencowe, D. Arneson, J. Ding, Y.-W. Chen, Z. Saleem, and X. Yang. Network modeling of single-cell omics data: challenges, opportunities, and progresses. *Emerging Topics in Life Sciences*, page ETLS20180176, jul 2019. ISSN 2397-8554. doi: 10.1042/{ETLS20180176}. URL http://www.emergtoplifesci.org/lookup/doi/10.1042/{ETLS20180176}.

D. Boehringer, E. M. Makarov, B. Sander, O. V. Makarova, B. Kastner, R. Lührmann, and H. Stark. Three-dimensional structure of a pre-catalytic human spliceosomal complex b. *Nature Structural & Molecular Biology*, 11(5):463–468, may 2004. doi: 10.1038/nsmb761. URL http://dx.doi.org/10.1038/nsmb761.

N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527, apr 2016. ISSN 1087-0156. doi: 10.1038/nbt.3519. URL http://www.nature.com/doifinder/10.1038/nbt.3519.

R. Breathnach, C. Benoist, K. O'Hare, F. Gannon, and P. Chambon. Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries. *Proceedings of the National Academy of Sciences of the United States of America*, 75(10):4853–4857, oct 1978. doi: 10.1073/pnas.75.10.4853. URL http://dx.doi.org/10.1073/pnas.75.10.4853.

F. Buettner, K. N. Natarajan, F. P. Casale, V. Proserpio, A. Scialdone, F. J. Theis, S. A. Teichmann, J. C. Marioni, and O. Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, 33(2):155–160, feb 2015. doi: 10.1038/nbt.3102. URL http://dx.doi.org/10.1038/nbt.3102.

I. Chambers, J. Silva, D. Colby, J. Nichols, B. Nijmeijer, M. Robertson, J. Vrana, K. Jones, L. Grotewold, and A. Smith. Nanog safeguards pluripotency and mediates germline development. *Nature*, 450(7173):1230–1234, dec 2007. ISSN 1476-4687. doi: 10.1038/nature06403. URL http://dx.doi.org/10.1038/nature06403.

S. Chen and J. C. Mar. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics*, 19(1):232, jun 2018. doi: 10.1186/s12859-018-2217-z. URL http://dx.doi.org/10.1186/s12859-018-2217-z.

W. Chen, H. P. Shulha, A. Ashar-Patel, J. Yan, K. M. Green, C. C. Query, N. Rhind, Z. Weng, and M. J. Moore. Endogenous u2·u5·u6 snRNA complexes in s. pombe are intron lariat spliceosomes. *RNA (New York)*, 20(3):308–320, mar 2014. doi: 10.1261/rna.040980.113. URL http://dx.doi.org/10.1261/rna.040980.113.

S. C. Cheng and J. Abelson. Spliceosome assembly in yeast. *Genes & Development*, 1(9):1014–1027, nov 1987. doi: 10.1101/gad.1.9.1014. URL http://dx.doi.org/10.1101/gad.1.9.1014.

L. T. Chow, R. E. Gelinas, T. R. Broker, and R. J. Roberts. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell*, 12(1):1–8, sep 1977. doi: 10.1016/0092-8674(77)90180-5. URL http://dx.doi.org/10.1016/0092-8674(77)90180-5.

C. Ciolli Mattioli, A. Rom, V. Franke, K. Imami, G. Arrey, M. Terne, A. Woehler, A. Akalin, I. Ulitsky, and M. Chekulaeva. Alternative 3' UTRs direct localization of functionally diverse protein isoforms in neuronal compartments. *Nucleic Acids Research*, 47(5):2560–2573, mar 2019. ISSN 0305-1048. doi: 10.1093/nar/gky1270. URL https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gky1270/5258023.

A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, and A. Mortazavi. A survey

of best practices for RNA-seq data analysis. *Genome Biology*, 17(1):13, jan 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-0881-8. URL http://genomebiology.com/2016/17/1/13.

J. Costa-Silva, D. Domingues, and F. M. Lopes. RNA-seq differential expression analysis: An extended review and a software tool. *Plos One*, 12(12):e0190152, dec 2017. doi: 10.1371/journal.pone.0190152. URL http://dx.doi.org/10.1371/journal.pone.0190152.

Z. Cvačková, D. Matějů, and D. Staněk. Retinitis pigmentosa mutations of SNRNP200 enhance cryptic splice-site recognition. *Human Mutation*, 35(3):308–317, mar 2014. doi: 10.1002/humu.22481. URL http://dx.doi.org/10.1002/humu.22481.

C. J. David and J. L. Manley. Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. *Genes & Development*, 24(21):2343–2364, nov 2010. doi: 10.1101/gad.1973010. URL http://dx.doi.org/10.1101/gad.1973010.

J. Deckert, K. Hartmuth, D. Boehringer, N. Behzadnia, C. L. Will, B. Kastner, H. Stark, H. Urlaub, and R. Lührmann. Protein composition and electron microscopy structure of affinity-purified human spliceosomal b complexes isolated under physiological conditions. *Molecular and Cellular Biology*, 26(14):5528–5543, jul 2006. doi: 10.1128/{MCB}.00582-06. URL http://dx.doi.org/10.1128/{MCB}.00582-06.

E. A. DePasquale, D. J. Schnell, I. Valiente, B. C. Blaxall, H. L. Grimes, H. Singh, and N. Salomonis. DoubletDecon: Cell-state aware removal of single-cell RNA-seq doublets. *BioRxiv*, jul 2018. doi: 10.1101/364810. URL http://biorxiv.org/lookup/doi/10.1101/364810.

A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. STAR: ultrafast universal RNA-seq aligner.

*Bioinformatics*, 29(1):15–21, jan 2013. doi: 10.1093/bioinformatics/bts635. URL http://dx.doi.org/10.1093/bioinformatics/bts635.

H. Du and M. Rosbash. The u1 snRNP protein U1C recognizes the 5' splice site in the absence of base pairing. *Nature*, 419(6902):86–90, sep 2002. ISSN 0028-0836. doi: 10.1038/nature00947. URL http://dx.doi.org/10.1038/nature00947.

S. Efroni, R. Duttagupta, J. Cheng, H. Dehghani, D. J. Hoeppner, C. Dash, D. P. Bazett-Jones, S. Le Grice, R. D. G. McKay, K. H. Buetow, T. R. Gingeras, T. Misteli, and E. Meshorer. Global transcription in pluripotent embryonic stem cells. *Cell Stem Cell*, 2(5):437–447, may 2008. ISSN 1934-5909. doi: 10.1016/j.stem.2008.03.021. URL http://dx.doi.org/10.1016/j.stem.2008.03.021.

S. J. Emrich, W. B. Barbazuk, L. Li, and P. S. Schnable. Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Research*, 17(1): 69–73, jan 2007. doi: 10.1101/gr.5145806. URL http://dx.doi.org/10.1101/gr.5145806.

C.-H. L. Eng, M. Lawson, Q. Zhu, R. Dries, N. Koulena, Y. Takei, J. Yun, C. Cronin, C. Karp, G.-C. Yuan, and L. Cai. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature*, 568(7751):235–239, apr 2019. ISSN 0028-0836. doi: 10.1038/s41586-019-1049-y. URL http://www.nature.com/articles/s41586-019-1049-y.

G. Eraslan, L. M. Simon, M. Mircea, N. S. Mueller, and F. J. Theis. Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications*, 10(1):390, jan 2019. ISSN 2041-1723. doi: 10.1038/s41467-018-07931-2. URL http://www.nature.com/articles/s41467-018-07931-2.

C. Everaert, M. Luypaert, J. L. V. Maag, Q. X. Cheng, M. E. Dinger, J. Hellemans, and P. Mestdagh. Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data. *Scientific Reports*, 7(1): 1559, may 2017. doi: 10.1038/s41598-017-01617-3. URL http://dx.doi.org/10.1038/s41598-017-01617-3.

P. Fabrizio, J. Dannenberg, P. Dube, B. Kastner, H. Stark, H. Urlaub, and R. Lührmann. The evolutionarily conserved core design of the catalytic activation step of the yeast spliceosome. *Molecular Cell*, 36(4):593–608, nov 2009. doi: 10.1016/j.molcel.2009.09.040. URL http://dx.doi.org/10.1016/j.molcel.2009.09.040.

T. Fei, T. Zhang, W. Shi, and T. Yu. Mitigating the adverse impact of batch effects in sample pattern detection. *Bioinformatics*, 34(15):2634–2641, aug 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty117. URL https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty117/4916062.

A. M. Femino, F. S. Fay, K. Fogarty, and R. H. Singer. Visualization of single RNA transcripts in situ. *Science*, 280(5363):585–590, apr 1998. doi: 10.1126/science.280.5363.585. URL http://dx.doi.org/10.1126/science.280.5363.585.

S. M. Fica, C. Oubridge, W. P. Galej, M. E. Wilkinson, X.-C. Bai, A. J. Newman, and K. Nagai. Structure of a spliceosome remodelled for exon ligation. *Nature*, 542(7641):377–380, feb 2017. doi: 10.1038/nature21078. URL http://dx.doi.org/10.1038/nature21078.

F. Finotello and B. Di Camillo. Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. *Briefings in functional genomics*, 14(2):130–142, mar 2015. doi: 10.1093/bfgp/elu035. URL http://dx.doi.org/10.1093/bfgp/elu035.

A. Frankish, M. Diekhans, A.-M. Ferreira, R. Johnson, I. Jungreis, J. Loveland, J. M. Mudge, C. Sisu, J. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, S. Carbonell Sala, J. Chrast, F. Cunningham, T. Di Domenico, S. Donaldson, I. T. Fiddes, C. García Girón, J. M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, T. Hunt, O. G. Izuogu, J. Lagarde, F. J. Martin, L. Martínez, S. Mohanan, P. Muir, F. C. P. Navarro, A. Parker, B. Pei, F. Pozo, M. Ruffier, B. M. Schmitt, E. Stapleton, M.-M. Suner, I. Sycheva, B. Uszczynska-Ratajczak, J. Xu,

A. Yates, D. Zerbino, Y. Zhang, B. Aken, J. S. Choudhary, M. Gerstein, R. Guigó, T. J. P. Hubbard, M. Kellis, B. Paten, A. Reymond, M. L. Tress, and P. Flicek. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, 47(D1):D766–D773, jan 2019. doi: 10.1093/nar/gky955. URL http://dx.doi.org/10.1093/nar/gky955.

A. C. Frazee, A. E. Jaffe, B. Langmead, and J. T. Leek. Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics*, 31(17): 2778–2784, sep 2015. doi: 10.1093/bioinformatics/btv272. URL http://dx.doi.org/10.1093/bioinformatics/btv272.

S. Fu, A. Wang, and K. F. Au. A comparative evaluation of hybrid error correction methods for error-prone long reads. *Genome Biology*, 20(1):26, feb 2019. ISSN 1474-760X. doi: 10.1186/s13059-018-1605-z. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1605-z.

E. Furman and D. G. Glitz. Purification of the spliceosome a-complex and its visualization by electron microscopy. *The Journal of Biological Chemistry*, 270(26): 15515–15522, jun 1995. doi: 10.1074/jbc.270.26.15515. URL http://dx.doi.org/10.1074/jbc.270.26.15515.

M. Gabut, P. Samavarchi-Tehrani, X. Wang, V. Slobodeniuc, D. O'Hanlon, H.-K. Sung, M. Alvarez, S. Talukder, Q. Pan, E. O. Mazzoni, S. Nedelec, H. Wichterle, K. Woltjen, T. R. Hughes, P. W. Zandstra, A. Nagy, J. L. Wrana, and B. J. Blencowe. An alternative splicing switch regulates embryonic stem cell pluripotency and reprogramming. *Cell*, 147(1):132–146, sep 2011. ISSN 1097-4172. doi: 10.1016/j.cell.2011.08.023. URL http://dx.doi.org/10.1016/j.cell.2011.08.023.

W. P. Galej, M. E. Wilkinson, S. M. Fica, C. Oubridge, A. J. Newman, and K. Nagai. Cryo-EM structure of the spliceosome immediately after branching. *Nature*, 537 (7619):197–201, sep 2016. doi: 10.1038/nature19316. URL http://dx.doi.org/10.1038/nature19316.

M. Garber, M. G. Grabherr, M. Guttman, and C. Trapnell. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*, 8(6):469–477, jun 2011. doi: 10.1038/nmeth.1613. URL http://dx.doi.org/10.1038/nmeth.1613.

P.-L. Germain, A. Vitriolo, A. Adamo, P. Laise, V. Das, and G. Testa. RNAontheBENCH: computational and empirical resources for benchmarking RNAseq quantification and differential expression methods. *Nucleic Acids Research*, 44 (11):5054–5067, jun 2016. doi: 10.1093/nar/gkw448. URL http://dx.doi.org/10.1093/nar/gkw448.

T. M. Gierahn, M. H. Wadsworth, T. K. Hughes, B. D. Bryson, A. Butler, R. Satija, S. Fortune, J. C. Love, and A. K. Shalek. Seq-well: portable, low-cost RNA sequencing of single cells at high throughput. *Nature Methods*, 14(4):395–398, apr 2017. doi: 10.1038/nmeth.4179. URL http://dx.doi.org/10.1038/nmeth.4179.

M. M. Golas, B. Sander, S. Bessonov, M. Grote, E. Wolf, B. Kastner, H. Stark, and R. Lührmann. 3D cryo-EM structure of an active step i spliceosome and localization of its catalytic core. *Molecular Cell*, 40(6):927–938, dec 2010. doi: 10.1016/j.molcel.2010.11.023. URL http://dx.doi.org/10.1016/j.molcel.2010.11.023.

W. Gong, I.-Y. Kwak, P. Pota, N. Koyano-Nakagawa, and D. J. Garry. DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics*, 19(1):220, jun 2018. doi: 10.1186/s12859-018-2226-y. URL http://dx.doi.org/10.1186/s12859-018-2226-y.

M. Gonzàlez-Porta, A. Frankish, J. Rung, J. Harrow, and A. Brazma. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biology*, 14(7):R70, jul 2013. doi: 10.1186/gb-2013-14-7-r70. URL http://dx.doi.org/10.1186/gb-2013-14-7-r70.

P. J. Grabowski and P. A. Sharp. Affinity chromatography of splicing complexes: U2, u5, and u4 + u6 small nuclear ribonucleoprotein particles in the spliceosome.

*Science*, 233(4770):1294–1299, sep 1986. doi: 10.1126/science.3638792. URL http://dx.doi.org/10.1126/science.3638792.

P. J. Grabowski, R. A. Padgett, and P. A. Sharp. Messenger RNA splicing in vitro: an excised intervening sequence and a potential intermediate. *Cell*, 37(2):415–427, jun 1984. doi: 10.1016/0092-8674(84)90372-6. URL http://dx.doi.org/10.1016/0092-8674(84)90372-6.

R. V. Grindberg, J. L. Yee-Greenbaum, M. J. McConnell, M. Novotny, A. L. O'Shaughnessy, G. M. Lambert, M. J. Araúzo-Bravo, J. Lee, M. Fishman, G. E. Robbins, X. Lin, P. Venepally, J. H. Badger, D. W. Galbraith, F. H. Gage, and R. S. Lasken. RNA-sequencing from single nuclei. *Proceedings of the National Academy of Sciences of the United States of America*, 110(49):19802–19807, dec 2013. doi: 10.1073/pnas.1319700110. URL http://dx.doi.org/10.1073/pnas.1319700110.

D. Grün and A. van Oudenaarden. Design and analysis of single-cell sequencing experiments. *Cell*, 163(4):799–810, nov 2015. ISSN 00928674. doi: 10.1016/j.cell.2015.10.039. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867415013537.

G. Guo, M. Huss, G. Q. Tong, C. Wang, L. Li Sun, N. D. Clarke, and P. Robson. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Developmental Cell*, 18(4):675–685, apr 2010. doi: 10.1016/j.devcel.2010.02.012. URL http://dx.doi.org/10.1016/j.devcel.2010.02.012.

I. Gupta, P. G. Collier, B. Haase, A. Mahfouz, A. Joglekar, T. Floyd, F. Koopmans, B. Barres, A. B. Smit, S. A. Sloan, W. Luo, O. Fedrigo, M. E. Ross, and H. U. Tilgner. Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nature Biotechnology*, oct 2018. ISSN 1087-0156. doi: 10.1038/nbt.4259. URL http://www.nature.com/doifinder/10.1038/nbt.4259.

J. Görnemann, C. Barrandon, K. Hujer, B. Rutz, G. Rigaut, K. M. Kotovic, C. Faux, K. M. Neugebauer, and B. Séraphin. Cotranscriptional spliceosome assembly and

splicing are independent of the prp40p WW domain. *RNA (New York)*, 17(12): 2119–2129, dec 2011. doi: 10.1261/rna.02646811. URL http://dx.doi.org/10.1261/rna.02646811.

M. Hagemann-Jensen, C. Ziegenhain, P. Chen, D. Ramskold, G.-J. Hendriks, A. J. Larsson, O. R. Faridani, and R. Sandberg. Single-cell RNA counting at allele- and isoform-resolution using smart-seq3. *BioRxiv*, oct 2019. doi: 10.1101/817924. URL http://biorxiv.org/lookup/doi/10.1101/817924.

L. Haghverdi, A. T. L. Lun, M. D. Morgan, and J. C. Marioni. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427, apr 2018. ISSN 1087-0156. doi: 10.1038/nbt.4091. URL http://www.nature.com/doifinder/10.1038/nbt.4091.

G. Haimovich and J. Gerst. Single-molecule fluorescence in situ hybridization (sm-FISH) for RNA detection in adherent animal cells. *Bio-protocol*, 8(21), 2018. ISSN 2331-8325. doi: 10.21769/{BioProtoc}.3070. URL https://bio-protocol.org/e3070.

J. Hang, R. Wan, C. Yan, and Y. Shi. Structural basis of pre-mRNA splicing. *Science*, 349(6253):1191–1198, sep 2015. ISSN 0036-8075. doi: 10.1126/science.aac8159. URL http://www.sciencemag.org/cgi/doi/10.1126/science.aac8159.

A. Haque, J. Engel, S. A. Teichmann, and T. Lönnberg. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine*, 9(1):75, aug 2017. doi: 10.1186/s13073-017-0467-4. URL http://dx.doi.org/10.1186/s13073-017-0467-4.

S. F. Hardy, P. J. Grabowski, R. A. Padgett, and P. A. Sharp. Cofactor requirements of splicing of purified messenger RNA precursors. *Nature*, 308(5957):375–377, mar 1984. doi: 10.1038/308375a0. URL http://dx.doi.org/10.1038/308375a0.

D. Haselbach, I. Komarov, D. E. Agafonov, K. Hartmuth, B. Graf, O. Dybkov, H. Urlaub, B. Kastner, R. Lührmann, and H. Stark. Structure and conformational dynamics of the human spliceosomal bact complex. *Cell*, 172(3):454–464.e11,

jan 2018. doi: 10.1016/j.cell.2018.01.010. URL http://dx.doi.org/10.1016/j.cell.2018.01.010.

T. Hashimshony, F. Wagner, N. Sher, and I. Yanai. CEL-seq: single-cell RNA-seq by multiplexed linear amplification. *Cell reports*, 2(3):666–673, sep 2012. doi: 10.1016/j.celrep.2012.08.003. URL http://dx.doi.org/10.1016/j.celrep.2012.08.003.

T. Hashimshony, N. Senderovich, G. Avital, A. Klochendler, Y. de Leeuw, L. Anavy, D. Gennert, S. Li, K. J. Livak, O. Rozenblatt-Rosen, Y. Dor, A. Regev, and I. Yanai. CEL-seq2: sensitive highly-multiplexed single-cell RNA-seq. *Genome Biology*, 17:77, apr 2016. doi: 10.1186/s13059-016-0938-8. URL http://dx.doi.org/10.1186/s13059-016-0938-8.

H. Henderson. Modern mathematicians. *Facts on File*, 1995.

N. Hernandez and W. Keller. Splicing of in vitro synthesized messenger RNA precursors in HeLa cell extracts. *Cell*, 35(1):89–99, nov 1983. doi: 10.1016/0092-8674(83)90211-8. URL http://dx.doi.org/10.1016/0092-8674(83)90211-8.

D. Hodgkins. Moments of discovery. *Kristallografiya*, 5(26). ISSN 1029-45.

J. Hu, E. Boritz, W. Wylie, and D. C. Douek. Stochastic principles governing alternative splicing of RNA. *PLoS Computational Biology*, 13(9):e1005761, sep 2017. doi: 10.1371/journal.pcbi.1005761. URL http://dx.doi.org/10.1371/journal.pcbi.1005761.

M. Huang, J. Wang, E. Torre, H. Dueck, S. Shaffer, R. Bonasio, J. I. Murray, A. Raj, M. Li, and N. R. Zhang. SAVER: gene expression recovery for single-cell RNA sequencing. *Nature Methods*, 15(7):539–542, jun 2018. ISSN 1548-7091. doi: 10.1038/s41592-018-0033-z. URL http://www.nature.com/articles/s41592-018-0033-z.

Y. Huang and G. Sanguinetti. BRIE: transcriptome-wide splicing quantification in single cells. *Genome Biology*, 18(1):123, jun 2017. doi: 10.1186/s13059-017-1248-5. URL http://dx.doi.org/10.1186/s13059-017-1248-5.

M. Hölzer and M. Marz. De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-seq assemblers. *GigaScience*, 8(5), may 2019. doi: 10.1093/gigascience/giz039. URL http://dx.doi.org/10.1093/gigascience/giz039.

J. O. Ilagan, R. J. Chalkley, A. L. Burlingame, and M. S. Jurica. Rearrangements within human spliceosomes captured after exon ligation. *RNA (New York)*, 19(3): 400–412, mar 2013. doi: 10.1261/rna.034223.112. URL http://dx.doi.org/10.1261/rna.034223.112.

T. Ilicic, J. K. Kim, A. A. Kolodziejczyk, F. O. Bagger, D. J. McCarthy, J. C. Marioni, and S. A. Teichmann. Classification of low quality cells from single-cell RNA-seq data. *Genome Biology*, 17:29, feb 2016. doi: 10.1186/s13059-016-0888-1. URL http://dx.doi.org/10.1186/s13059-016-0888-1.

S. Islam, U. Kjällquist, A. Moliner, P. Zajac, J.-B. Fan, P. Lönnerberg, and S. Linnarsson. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Research*, 21(7):1160–1167, jul 2011. ISSN 1549-5469. doi: 10.1101/gr.110882.110. URL http://dx.doi.org/10.1101/gr.110882.110.

S. Islam, A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. Lönnerberg, and S. Linnarsson. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, 11(2):163–166, feb 2014. doi: 10.1038/nmeth.2772. URL http://dx.doi.org/10.1038/nmeth.2772.

D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay, and I. Amit. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*, 343(6172):776–779, feb 2014. doi: 10.1126/science.1247651. URL http://dx.doi.org/10.1126/science.1247651.

J. L. Jenkins, A. A. Agrawal, A. Gupta, M. R. Green, and C. L. Kielkopf. U2AF65 adapts to diverse pre-mRNA splice sites through conformational selection of specific and promiscuous RNA recognition motifs. *Nucleic Acids Research*, 41(6): 3859–3873, apr 2013. doi: 10.1093/nar/gkt046. URL http://dx.doi.org/10.1093/nar/gkt046.

L. Jiang, F. Schlesinger, C. A. Davis, Y. Zhang, R. Li, M. Salit, T. R. Gingeras, and B. Oliver. Synthetic spike-in standards for RNA-seq experiments. *Genome Research*, 21(9):1543–1551, sep 2011. doi: 10.1101/gr.121095.111. URL http://dx.doi.org/10.1101/gr.121095.111.

M. S. Jurica, D. Sousa, M. J. Moore, and N. Grigorieff. Three-dimensional structure of c complex spliceosomes by electron microscopy. *Nature Structural & Molecular Biology*, 11(3):265–269, mar 2004. ISSN 1545-9993. doi: 10.1038/nsmb728. URL http://dx.doi.org/10.1038/nsmb728.

T. Kanagawa. Bias and artifacts in multitemplate polymerase chain reactions (PCR). *Journal of Bioscience and Bioengineering*, 96(4):317–323, 2003. doi: 10.1016/S1389-1723(03)90130-7. URL http://dx.doi.org/10.1016/S1389-1723(03)90130-7.

H. M. Kang, M. Subramaniam, S. Targ, M. Nguyen, L. Maliskova, E. McCarthy, E. Wan, S. Wong, L. Byrnes, C. M. Lanata, R. E. Gate, S. Mostafavi, A. Marson, N. Zaitlen, L. A. Criswell, and C. J. Ye. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology*, 36(1):89–94, 2018. ISSN 1087-0156. doi: 10.1038/nbt.4042. URL http://www.nature.com/doifinder/10.1038/nbt.4042.

K. Karlsson and S. Linnarsson. Single-cell mRNA isoform diversity in the mouse brain. *BMC Genomics*, 18(1):126, feb 2017. doi: 10.1186/s12864-017-3528-6. URL http://dx.doi.org/10.1186/s12864-017-3528-6.

Y. Katz, E. T. Wang, E. M. Airoldi, and C. B. Burge. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, 7(12):

1009–1015, dec 2010. doi: 10.1038/nmeth.1528. URL http://dx.doi.org/10.1038/nmeth.1528.

P. V. Kharchenko, L. Silberstein, and D. T. Scadden. Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11(7):740–742, jul 2014. doi: 10.1038/nmeth.2967. URL http://dx.doi.org/10.1038/nmeth.2967.

J. K. Kim and J. C. Marioni. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biology*, 14(1):R7, jan 2013. doi: 10.1186/gb-2013-14-1-r7. URL http://dx.doi.org/10.1186/gb-2013-14-1-r7.

J. K. Kim, A. A. Kolodziejczyk, T. Ilicic, S. A. Teichmann, and J. C. Marioni. Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nature Communications*, 6:8687, oct 2015. doi: 10.1038/ncomms9687. URL http://dx.doi.org/10.1038/ncomms9687.

A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, may 2015. doi: 10.1016/j.cell.2015.04.044. URL http://dx.doi.org/10.1016/j.cell.2015.04.044.

A. A. Kolodziejczyk, J. K. Kim, J. C. H. Tsang, T. Ilicic, J. Henriksson, K. N. Natarajan, A. C. Tuck, X. Gao, M. Bühler, P. Liu, J. C. Marioni, and S. A. Teichmann. Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell*, 17(4):471–485, oct 2015. doi: 10.1016/j.stem.2015.09.011. URL http://dx.doi.org/10.1016/j.stem.2015.09.011.

M. M. Konarska and P. A. Sharp. Electrophoretic separation of complexes involved in the splicing of precursors to mRNAs. *Cell*, 46(6):845–855, sep 1986. doi: 10.1016/0092-8674(86)90066-8. URL http://dx.doi.org/10.1016/0092-8674(86)90066-8.

M. M. Konarska and P. A. Sharp. Interactions between small nuclear ribonucleoprotein particles in formation of spliceosomes. *Cell*, 49(6):763–774, jun

1987. doi: 10.1016/0092-8674(87)90614-3. URL http://dx.doi.org/10.1016/0092-8674(87)90614-3.

Y. Kondo, C. Oubridge, A.-M. M. van Roon, and K. Nagai. Crystal structure of human u1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5' splice site recognition. *eLife*, 4, jan 2015. doi: 10.7554/{eLife}.04986. URL http://dx.doi.org/10.7554/{eLife}.04986.

S. Koren, M. C. Schatz, B. P. Walenz, J. Martin, J. T. Howard, G. Ganapathy, Z. Wang, D. A. Rasko, W. R. McCombie, E. D. Jarvis, and A. M. Phillippy. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology*, 30(7):693–700, jul 2012. doi: 10.1038/nbt.2280. URL http://dx.doi.org/10.1038/nbt.2280.

A. R. Krainer and T. Maniatis. Multiple factors including the small nuclear ribonucleoproteins u1 and u2 are necessary for pre-mRNA splicing in vitro. *Cell*, 42(3): 725–736, oct 1985. doi: 10.1016/0092-8674(85)90269-7. URL http://dx.doi.org/10.1016/0092-8674(85)90269-7.

A. Lafzi, C. Moutinho, S. Picelli, and H. Heyn. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. *Nature Protocols*, 13 (12):2742–2757, 2018. ISSN 1754-2189. doi: 10.1038/s41596-018-0073-y. URL http://www.nature.com/articles/s41596-018-0073-y.

A. I. Lamond, M. M. Konarska, and P. A. Sharp. A mutational analysis of spliceosome assembly: evidence for splice site collaboration during spliceosome formation. *Genes & Development*, 1(6):532–543, aug 1987. doi: 10.1101/gad.1.6.532. URL http://dx.doi.org/10.1101/gad.1.6.532.

A. J. M. Larsson, P. Johnsson, M. Hagemann-Jensen, L. Hartmanis, O. R. Faridani, B. Reinius, r. Segerstolpe, C. M. Rivera, B. Ren, and R. Sandberg. Genomic encoding of transcriptional burst kinetics. *Nature*, 565(7738):251–254, jan 2019. ISSN 0028-0836. doi: 10.1038/s41586-018-0836-1. URL http://www.nature.com/articles/s41586-018-0836-1.

K. Lebrigand, V. Magnone, P. Barbry, and R. Waldmann. High throughput, error corrected nanopore single cell transcriptome sequencing. *BioRxiv*, nov 2019. doi: 10.1101/831495. URL http://biorxiv.org/lookup/doi/10.1101/831495.

M. R. Lerner and J. A. Steitz. Antibodies to small nuclear RNAs complexed with proteins are produced by patients with systemic lupus erythematosus. *Proceedings of the National Academy of Sciences of the United States of America*, 76(11): 5495–5499, nov 1979. doi: 10.1073/pnas.76.11.5495. URL http://dx.doi.org/10.1073/pnas.76.11.5495.

M. R. Lerner, J. A. Boyle, S. M. Mount, S. L. Wolin, and J. A. Steitz. Are snRNPs involved in splicing? *Nature*, 283(5743):220–224, jan 1980. doi: 10.1038/283220a0. URL http://dx.doi.org/10.1038/283220a0.

C. F. Lesser and C. Guthrie. Mutations in u6 snRNA that alter splice site specificity: implications for the active site. *Science*, 262(5142):1982–1988, dec 1993. doi: 10.1126/science.8266093. URL http://dx.doi.org/10.1126/science.8266093.

A. K. W. Leung, K. Nagai, and J. Li. Structure of the spliceosomal u4 snRNP core domain and its implication for snRNP biogenesis. *Nature*, 473(7348):536–539, may 2011. doi: 10.1038/nature09956. URL http://dx.doi.org/10.1038/nature09956.

M. J. Levesque, P. Ginart, Y. Wei, and A. Raj. Visualizing SNVs to quantify allele-specific expression in single cells. *Nature Methods*, 10(9):865–867, sep 2013. doi: 10.1038/nmeth.2589. URL http://dx.doi.org/10.1038/nmeth.2589.

B. Li and C. N. Dewey. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics*, 12:323, aug 2011. doi: 10.1186/1471-2105-12-323. URL http://dx.doi.org/10.1186/1471-2105-12-323.

H. Li. lh3/seqtk: Toolkit for processing sequences in FASTA/q formats, 2013. URL https://github.com/lh3/seqtk.

J. Li, Y. Wang, X. Rao, Y. Wang, W. Feng, H. Liang, and Y. Liu. Roles of alternative splicing in modulating transcriptional regulation. *BMC Systems Biology*, 11(Suppl 5):89, oct 2017. doi: 10.1186/s12918-017-0465-6. URL http://dx.doi.org/10.1186/s12918-017-0465-6.

W. V. Li and J. J. Li. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nature Communications*, 9(1):997, mar 2018. doi: 10.1038/s41467-018-03405-7. URL http://dx.doi.org/10.1038/s41467-018-03405-7.

K. H. Lim, L. Ferraris, M. E. Filloux, B. J. Raphael, and W. G. Fairbrother. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proceedings of the National Academy of Sciences of the United States of America*, 108(27):11093–11098, jul 2011. doi: 10.1073/pnas.1101135108. URL http://dx.doi.org/10.1073/pnas.1101135108.

P.-C. Lin and R.-M. Xu. Structure and assembly of the SF3a splicing factor complex of u2 snRNP. *The EMBO Journal*, 31(6):1579–1590, mar 2012. doi: 10.1038/emboj.2012.7. URL http://dx.doi.org/10.1038/emboj.2012.7.

R. Lister, R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, A. H. Millar, and J. R. Ecker. Highly integrated single-base resolution maps of the epigenome in arabidopsis. *Cell*, 133(3):523–536, may 2008. ISSN 1097-4172. doi: 10.1016/j.cell.2008.03.029. URL http://dx.doi.org/10.1016/j.cell.2008.03.029.

M. D. Luecken and F. J. Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746, jun 2019. ISSN 1744-4292. doi: 10.15252/msb.20188746. URL http://msb.embopress.org/lookup/doi/10.15252/msb.20188746.

D. Lukacsovich, J. Winterer, L. Que, W. Luo, T. Lukacsovich, and C. Földy. Single-cell RNA-seq reveals developmental origins and ontogenetic stability of neurexin alternative splicing profiles. *Cell reports*, 27(13):3752–3759.e4, jun 2019. ISSN 22111247. doi: 10.1016/j.celrep.2019.05.090. URL https://linkinghub.elsevier.com/retrieve/pii/S2211124719307296.

A. T. L. Lun and J. C. Marioni. Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data. *Biostatistics*, 18(3):451–464, jul 2017. doi: 10.1093/biostatistics/kxw055. URL http://dx.doi.org/10.1093/biostatistics/kxw055.

N. López-Bigas, B. Audit, C. Ouzounis, G. Parra, and R. Guigó. Are splicing mutations the most frequent cause of hereditary disease? *FEBS Letters*, 579(9):1900–1903, mar 2005. doi: 10.1016/j.febslet.2005.02.047. URL http://dx.doi.org/10.1016/j.febslet.2005.02.047.

E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161 (5):1202–1214, may 2015. ISSN 00928674. doi: 10.1016/j.cell.2015.05.002. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867415005498.

H. D. Madhani and C. Guthrie. A novel base-pairing interaction between u2 and u6 snRNAs suggests a mechanism for the catalytic activation of the spliceosome. *Cell*, 71(5):803–817, nov 1992. doi: 10.1016/0092-8674(92)90556-r. URL http://dx.doi.org/10.1016/0092-8674(92)90556-r.

G. K. Marinov, B. A. Williams, K. McCue, G. P. Schroth, J. Gertz, R. M. Myers, and B. J. Wold. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Research*, 24(3):496–510, mar 2014. doi: 10.1101/gr.161034.113. URL http://dx.doi.org/10.1101/gr.161034.113.

H. Marks, T. Kalkan, R. Menafra, S. Denissov, K. Jones, H. Hofemeister, J. Nichols, A. Kranz, A. F. Stewart, A. Smith, and H. G. Stunnenberg. The transcriptional and epigenomic foundations of ground state pluripotency. *Cell*, 149(3):590–604, apr 2012. ISSN 00928674. doi: 10.1016/j.cell.2012.03.026. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867412004096.

M. Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10, may 2011. ISSN 2226-6089. doi: 10.14806/ej.17.1.200. URL http://journal.embnet.org/index.php/embnetjournal/article/view/200.

A. G. Matera and Z. Wang. A day in the life of the spliceosome. *Nature Reviews. Molecular Cell Biology*, 15(2):108–121, feb 2014. doi: 10.1038/nrm3742. URL http://dx.doi.org/10.1038/nrm3742.

D. J. McCarthy, K. R. Campbell, A. T. L. Lun, and Q. F. Wills. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in r. *Bioinformatics*, 33(8):1179–1186, apr 2017. doi: 10.1093/bioinformatics/btw777. URL http://dx.doi.org/10.1093/bioinformatics/btw777.

K. E. McElroy, F. Luciani, and T. Thomas. GemSIM: general, error-model based simulator of next-generation sequencing data. *BMC Genomics*, 13:74, feb 2012. doi: 10.1186/1471-2164-13-74. URL http://dx.doi.org/10.1186/1471-2164-13-74.

J. R. Moffitt, J. Hao, G. Wang, K. H. Chen, H. P. Babcock, and X. Zhuang. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proceedings of the National Academy of Sciences of the United States of America*, 113(39):11046–11051, sep 2016. doi: 10.1073/pnas.1612826113. URL http://dx.doi.org/10.1073/pnas.1612826113.

E. J. Montemayor, E. C. Curran, H. H. Liao, K. L. Andrews, C. N. Treba, S. E. Butcher, and D. A. Brow. Core structure of the u6 small nuclear ribonucleoprotein at 1.7-å resolution. *Nature Structural & Molecular Biology*, 21(6):544–551, jun 2014. doi: 10.1038/nsmb.2832. URL http://dx.doi.org/10.1038/nsmb.2832.

S. Morgani, J. Nichols, and A.-K. Hadjantonakis. The many faces of pluripotency: in vitro adaptations of a continuum of in vivo states. *BMC Developmental Biology*, 17(1):7, jun 2017. doi: 10.1186/s12861-017-0150-4. URL http://dx.doi.org/10.1186/s12861-017-0150-4.

D. P. Morris and A. L. Greenleaf. The splicing factor, prp40, binds the phosphorylated carboxyl-terminal domain of RNA polymerase II. *The Journal of Biological Chemistry*, 275(51):39935–39943, dec 2000. doi: 10.1074/jbc.M004118200. URL http://dx.doi.org/10.1074/jbc.M004118200.

S. M. Mount. A catalogue of splice junction sequences. *Nucleic Acids Research*, 10 (2):459–472, jan 1982. doi: 10.1093/nar/10.2.459. URL http://dx.doi.org/10.1093/nar/10.2.459.

S. M. Mount. Genomic sequence, splicing, and gene annotation. *American Journal of Human Genetics*, 67(4):788–792, oct 2000. doi: 10.1086/303098. URL http://dx.doi.org/10.1086/303098.

F. Muntoni, S. Torelli, and A. Ferlini. Dystrophin and mutations: one gene, several proteins, multiple phenotypes. *Lancet Neurology*, 2(12):731–740, dec 2003. doi: 10.1016/S1474-4422(03)00585-4. URL http://dx.doi.org/10.1016/S1474-4422(03)00585-4.

NatMethods. Method of the year 2013. *Nature Methods*, 11(1):1, jan 2014. doi: 10.1038/nmeth.2801. URL http://dx.doi.org/10.1038/nmeth.2801.

A. Newman and C. Norman. Mutations in yeast u5 snRNA alter the specificity of 5' splice-site cleavage. *Cell*, 65(1):115–123, apr 1991. doi: 10.1016/0092-8674(91)90413-s. URL http://dx.doi.org/10.1016/0092-8674(91)90413-s.

A. J. Newman, S. Teigelkamp, and J. D. Beggs. snRNA interactions at 5' and 3' splice sites monitored by photoactivated crosslinking in yeast spliceosomes. *RNA (New York)*, 1(9):968–980, nov 1995. URL https://www.ncbi.nlm.nih.gov/pubmed/8548661.

T. H. D. Nguyen, W. P. Galej, X.-C. Bai, C. Oubridge, A. J. Newman, S. H. W. Scheres, and K. Nagai. Cryo-EM structure of the yeast u4/u6.u5 tri-snRNP at 3.7 å resolution. *Nature*, 530(7590):298–302, feb 2016. doi: 10.1038/nature16940. URL http://dx.doi.org/10.1038/nature16940.

H. Ochiai, T. Hayashi, M. Umeda, M. Yoshimura, A. Harada, Y. Shimizu, K. Nakano, N. Saitoh, H. Kimura, Z. Liu, T. Yamamoto, T. Okamura, Y. Ohkawa, and I. Nikaido. Genome-wide analysis of transcriptional bursting-induced noise in mammalian cells. *BioRxiv*, aug 2019. doi: 10.1101/736207. URL `http://biorxiv.org/lookup/doi/10.1101/736207`.

M. D. Ohi, L. Ren, J. S. Wall, K. L. Gould, and T. Walz. Structural characterization of the fission yeast u5.u2/u6 spliceosome complex. *Proceedings of the National Academy of Sciences of the United States of America*, 104(9):3195–3200, feb 2007. doi: 10.1073/pnas.0611591104. URL `http://dx.doi.org/10.1073/pnas.0611591104`.

R. A. Padgett, M. M. Konarska, P. J. Grabowski, S. F. Hardy, and P. A. Sharp. Lariat RNAs as intermediates and products in the splicing of messenger RNA precursors. *Science*, 225(4665):898–903, aug 1984. doi: 10.1126/science.6206566. URL `http://dx.doi.org/10.1126/science.6206566`.

R. Patro, S. M. Mount, and C. Kingsford. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature Biotechnology*, 32(5):462–464, may 2014. doi: 10.1038/nbt.2862. URL `http://dx.doi.org/10.1038/nbt.2862`.

R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14 (4):417–419, apr 2017. doi: 10.1038/nmeth.4197. URL `http://dx.doi.org/10.1038/nmeth.4197`.

R. Petegrosso, Z. Li, and R. Kuang. Machine learning and statistical methods for clustering single-cell RNA-sequencing data. *Briefings in Bioinformatics*, jun 2019. doi: 10.1093/bib/bbz063. URL `http://dx.doi.org/10.1093/bib/bbz063`.

V. Petukhov, J. Guo, N. Baryawno, N. Severe, D. T. Scadden, M. G. Samsonova, and P. V. Kharchenko. dropEst: pipeline for accurate estimation of molecular counts in droplet-based single-cell RNA-seq experiments. *Genome Biology*, 19(1):

78, jun 2018. ISSN 1474-760X. doi: 10.1186/s13059-018-1449-6. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1449-6.

B. Phipson, L. Zappia, and A. Oshlack. Gene length and detection bias in single cell RNA sequencing protocols. *F1000Research*, 6:595, apr 2017. doi: 10.12688/f1000research.11290.1. URL http://dx.doi.org/10.12688/f1000research.11290.1.

S. Picelli. Single-cell RNA-sequencing: The future of genome biology is now. *RNA Biology*, 14(5):637–650, may 2017. doi: 10.1080/15476286.2016.1201618. URL http://dx.doi.org/10.1080/15476286.2016.1201618.

S. Picelli, O. R. Faridani, A. K. Björklund, G. Winberg, S. Sagasser, and R. Sandberg. Full-length RNA-seq from single cells using smart-seq2. *Nature Protocols*, 9(1):171–181, jan 2014. doi: 10.1038/nprot.2014.006. URL http://dx.doi.org/10.1038/nprot.2014.006.

C. W. Pikielny, B. C. Rymond, and M. Rosbash. Electrophoresis of ribonucleoproteins reveals an ordered assembly pathway of yeast splicing complexes. *Nature*, 324(6095):341–345, 1986. doi: 10.1038/324341a0. URL http://dx.doi.org/10.1038/324341a0.

D. A. Pomeranz Krummel, C. Oubridge, A. K. W. Leung, J. Li, and K. Nagai. Crystal structure of human spliceosomal u1 snRNP at 5.5 a resolution. *Nature*, 458(7237):475–480, mar 2009. ISSN 1476-4687. doi: 10.1038/nature07851. URL http://dx.doi.org/10.1038/nature07851.

S. R. Price, P. R. Evans, and K. Nagai. Crystal structure of the spliceosomal U2B"-U2A' protein complex bound to a fragment of u2 small nuclear RNA. *Nature*, 394(6694):645–650, aug 1998. doi: 10.1038/29234. URL http://dx.doi.org/10.1038/29234.

P. L. Raghunathan and C. Guthrie. RNA unwinding in u4/u6 snRNPs requires ATP hydrolysis and the DEIH-box splicing factor brr2. *Current Biology*, 8(15):

847–855, jul 1998. doi: 10.1016/s0960-9822(07)00345-4. URL http://dx.doi.org/10.1016/s0960-9822(07)00345-4.

D. Ramsköld, S. Luo, Y.-C. Wang, R. Li, Q. Deng, O. R. Faridani, G. A. Daniels, I. Khrebtukova, J. F. Loring, L. C. Laurent, G. P. Schroth, and R. Sandberg. Full-length mRNA-seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotechnology*, 30(8):777–782, aug 2012. doi: 10.1038/nbt.2282. URL http://dx.doi.org/10.1038/nbt.2282.

F. J. Rang, W. P. Kloosterman, and J. de Ridder. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biology*, 19(1):90, jul 2018. ISSN 1474-760X. doi: 10.1186/s13059-018-1462-9. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1462-9.

R. Rauhut, P. Fabrizio, O. Dybkov, K. Hartmuth, V. Pena, A. Chari, V. Kumar, C.-T. Lee, H. Urlaub, B. Kastner, H. Stark, and R. Lührmann. Molecular architecture of the saccharomyces cerevisiae activated spliceosome. *Science*, 353(6306):1399–1405, sep 2016. doi: 10.1126/science.aag1906. URL http://dx.doi.org/10.1126/science.aag1906.

M. G. Reese, G. Hartzell, N. L. Harris, U. Ohler, J. F. Abril, and S. E. Lewis. Genome annotation assessment in drosophila melanogaster. *Genome Research*, 10 (4):483–501, apr 2000. doi: 10.1101/gr.10.4.483. URL http://dx.doi.org/10.1101/gr.10.4.483.

A. Regev, S. A. Teichmann, E. S. Lander, I. Amit, C. Benoist, E. Birney, B. Bodenmiller, P. Campbell, P. Carninci, M. Clatworthy, H. Clevers, B. Deplancke, I. Dunham, J. Eberwine, R. Eils, W. Enard, A. Farmer, L. Fugger, B. Göttgens, N. Hacohen, M. Haniffa, M. Hemberg, S. Kim, P. Klenerman, A. Kriegstein, E. Lein, S. Linnarsson, E. Lundberg, J. Lundeberg, P. Majumder, J. C. Marioni, M. Merad, M. Mhlanga, M. Nawijn, M. Netea, G. Nolan, D. Pe'er, A. Phillipakis, C. P. Ponting, S. Quake, W. Reik, O. Rozenblatt-Rosen, J. Sanes, R. Satija, T. N.

Schumacher, A. Shalek, E. Shapiro, P. Sharma, J. W. Shin, O. Stegle, M. Stratton, M. J. T. Stubbington, F. J. Theis, M. Uhlen, A. van Oudenaarden, A. Wagner, F. Watt, J. Weissman, B. Wold, R. Xavier, N. Yosef, and H. C. A. M. Participants. The human cell atlas. *eLife*, 6, dec 2017. doi: 10.7554/{eLife}.27041. URL http://dx.doi.org/10.7554/{eLife}.27041.

A. Roberts and L. Pachter. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature Methods*, 10(1):71–73, jan 2013. doi: 10.1038/nmeth.2251. URL http://dx.doi.org/10.1038/nmeth.2251.

S. G. Rodriques, R. R. Stickels, A. Goeva, C. A. Martin, E. Murray, C. R. Vanderburg, J. Welch, L. M. Chen, F. Chen, and E. Z. Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467, mar 2019. ISSN 0036-8075. doi: 10.1126/science.aaw1219. URL http://www.sciencemag.org/lookup/doi/10.1126/science.aaw1219.

B. Ruskin, A. R. Krainer, T. Maniatis, and M. R. Green. Excision of an intact intron as a novel lariat structure during pre-mRNA splicing in vitro. *Cell*, 38(1):317–331, aug 1984. doi: 10.1016/0092-8674(84)90553-1. URL http://dx.doi.org/10.1016/0092-8674(84)90553-1.

W. Saelens, R. Cannoodt, H. Todorov, and Y. Saeys. A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, 37(5):547–554, apr 2019. ISSN 1087-0156. doi: 10.1038/s41587-019-0071-9. URL http://www.nature.com/articles/s41587-019-0071-9.

L. Samaranch, O. Lorenzo-Betancor, J. M. Arbelo, I. Ferrer, E. Lorenzo, J. Irigoyen, M. A. Pastor, C. Marrero, C. Isla, J. Herrera-Henriquez, and P. Pastor. PINK1-linked parkinsonism is associated with lewy body pathology. *Brain: A Journal of Neurology*, 133(Pt 4):1128–1142, apr 2010. ISSN 1460-2156. doi: 10.1093/brain/awq051. URL http://dx.doi.org/10.1093/brain/awq051.

A. K. Shalek, R. Satija, X. Adiconis, R. S. Gertner, J. T. Gaublomme, R. Raychowdhury, S. Schwartz, N. Yosef, C. Malboeuf, D. Lu, J. J. Trombetta, D. Gennert, A. Gnirke, A. Goren, N. Hacohen, J. Z. Levin, H. Park, and A. Regev. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, 498(7453):236–240, jun 2013. doi: 10.1038/nature12172. URL http://dx.doi.org/10.1038/nature12172.

K. Shekhar, S. W. Lapan, I. E. Whitney, N. M. Tran, E. Z. Macosko, M. Kowalczyk, X. Adiconis, J. Z. Levin, J. Nemesh, M. Goldman, S. A. McCarroll, C. L. Cepko, A. Regev, and J. R. Sanes. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell*, 166(5):1308–1323.e30, aug 2016. doi: 10.1016/j.cell.2016.07.054. URL http://dx.doi.org/10.1016/j.cell.2016.07.054.

Y. Shi. Mechanistic insights into precursor messenger RNA splicing by the spliceosome. *Nature Reviews. Molecular Cell Biology*, 18(11):655–670, nov 2017. doi: 10.1038/nrm.2017.86. URL http://dx.doi.org/10.1038/nrm.2017.86.

E. A. Sickmier, K. E. Frato, H. Shen, S. R. Paranawithana, M. R. Green, and C. L. Kielkopf. Structural basis for polypyrimidine tract recognition by the essential pre-mRNA splicing factor U2AF65. *Molecular Cell*, 23(1):49–59, jul 2006. doi: 10.1016/j.molcel.2006.05.025. URL http://dx.doi.org/10.1016/j.molcel.2006.05.025.

M. Singh, G. Al-Eryani, S. Carswell, J. M. Ferguson, J. Blackburn, K. Barton, D. Roden, F. Luciani, T. Giang Phan, S. Junankar, K. Jackson, C. C. Goodnow, M. A. Smith, and A. Swarbrick. High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. *Nature Communications*, 10(1):3120, jul 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-11049-4. URL http://www.nature.com/articles/s41467-019-11049-4.

T. Smith, A. Heger, and I. Sudbery. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Research*, 27(3):

491–499, jan 2017. doi: 10.1101/gr.209601.116. URL http://dx.doi.org/10.1101/gr.209601.116.

C. Soneson and M. D. Robinson. Towards unified quality verification of synthetic count data with countsimQC. *Bioinformatics*, 34(4):691–692, feb 2018a. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx631. URL http://academic.oup.com/bioinformatics/article/doi/10.1093/bioinformatics/btx631/4345646/Towards-unified-quality-verification-of-synthetic.

C. Soneson and M. D. Robinson. Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, 15(4):255–261, feb 2018b. ISSN 1548-7091. doi: 10.1038/nmeth.4612. URL http://www.nature.com/doifinder/10.1038/nmeth.4612.

Y. Song, O. B. Botvinnik, M. T. Lovci, B. Kakaradov, P. Liu, J. L. Xu, and G. W. Yeo. Single-cell alternative splicing analysis with expedition reveals splicing dynamics during neuron differentiation. *Molecular Cell*, 67(1):148–161.e5, jul 2017. doi: 10.1016/j.molcel.2017.06.003. URL http://dx.doi.org/10.1016/j.molcel.2017.06.003.

E. J. Sontheimer and J. A. Steitz. The u5 and u6 small nuclear RNAs as active site components of the spliceosome. *Science*, 262(5142):1989–1996, dec 1993. doi: 10.1126/science.8266094. URL http://dx.doi.org/10.1126/science.8266094.

O. Stegle, S. A. Teichmann, and J. C. Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews. Genetics*, 16(3):133–145, mar 2015. doi: 10.1038/nrg3833. URL http://dx.doi.org/10.1038/nrg3833.

T. Sterne-Weiler, J. Howard, M. Mort, D. N. Cooper, and J. R. Sanford. Loss of exon identity is a common mechanism of human inherited disease. *Genome Research*, 21(10):1563–1571, oct 2011. doi: 10.1101/gr.118638.110. URL http://dx.doi.org/10.1101/gr.118638.110.

M. Stoeckius, S. Zheng, B. Houck-Loomis, S. Hao, B. Z. Yeung, W. M. Mauck, P. Smibert, and R. Satija. Cell hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biology*, 19(1):224, dec 2018. doi: 10.1186/s13059-018-1603-1. URL `http://dx.doi.org/10.1186/s13059-018-1603-1`.

E. W. Strong. Newton's "mathematical way". *Journal of the history of ideas*, 12(1): 90, jan 1951. ISSN 00225037. doi: 10.2307/2707539. URL `https://www.jstor.org/stable/2707539?origin=crossref`.

P. L. Ståhl, F. Salmén, S. Vickovic, A. Lundmark, J. F. Navarro, J. Magnusson, S. Giacomello, M. Asp, J. O. Westholm, M. Huss, A. Mollbrink, S. Linnarsson, S. Codeluppi, r. Borg, F. Pontén, P. I. Costea, P. Sahlén, J. Mulder, O. Bergmann, J. Lundeberg, and J. Frisén. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, jul 2016. ISSN 0036-8075. doi: 10.1126/science.aaf2403. URL `http://www.sciencemag.org/cgi/doi/10.1126/science.aaf2403`.

C.-H. Su, D. D, and W.-Y. Tarn. Alternative splicing in neurogenesis and brain development. *Frontiers in molecular biosciences*, 5:12, feb 2018. doi: 10.3389/fmolb.2018.00012. URL `http://dx.doi.org/10.3389/fmolb.2018.00012`.

V. Svensson, K. N. Natarajan, L.-H. Ly, R. J. Miragaia, C. Labalette, I. C. Macaulay, A. Cvejic, and S. A. Teichmann. Power analysis of single-cell RNA-sequencing experiments. *Nature Methods*, 14(4):381–387, apr 2017. doi: 10.1038/nmeth.4220. URL `http://dx.doi.org/10.1038/nmeth.4220`.

V. Svensson, R. Vento-Tormo, and S. A. Teichmann. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols*, 13(4):599–604, mar 2018. ISSN 1754-2189. doi: 10.1038/nprot.2017.149. URL `http://www.nature.com/doifinder/10.1038/nprot.2017.149`.

V. Svensson, E. da Veiga Beltrame, and L. Pachter. Quantifying the tradeoff between

sequencing depth and cell number in single-cell RNA-seq. *BioRxiv*, sep 2019. doi: 10.1101/762773. URL http://biorxiv.org/lookup/doi/10.1101/762773.

G. Tanackovic, A. Ransijn, C. Ayuso, S. Harper, E. L. Berson, and C. Rivolta. A missense mutation in PRPF6 causes impairment of pre-mRNA splicing and autosomal-dominant retinitis pigmentosa. *American Journal of Human Genetics*, 88(5):643–649, may 2011. doi: 10.1016/j.ajhg.2011.04.008. URL http://dx.doi.org/10.1016/j.ajhg.2011.04.008.

F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao, and M. A. Surani. mRNA-seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382, may 2009. ISSN 1548-7105. doi: 10.1038/nmeth.1315. URL http://dx.doi.org/10.1038/nmeth.1315.

F. Tang, C. Barbacioru, S. Bao, C. Lee, E. Nordman, X. Wang, K. Lao, and M. A. Surani. Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-seq analysis. *Cell Stem Cell*, 6(5):468–478, may 2010. doi: 10.1016/j.stem.2010.03.015. URL http://dx.doi.org/10.1016/j.stem.2010.03.015.

S. Tarazona, F. García-Alcalde, J. Dopazo, A. Ferrer, and A. Conesa. Differential expression in RNA-seq: a matter of depth. *Genome Research*, 21(12):2213–2223, dec 2011. doi: 10.1101/gr.124321.111. URL http://dx.doi.org/10.1101/gr.124321.111.

M. Teng, M. I. Love, C. A. Davis, S. Djebali, A. Dobin, B. R. Graveley, S. Li, C. E. Mason, S. Olson, D. Pervouchine, C. A. Sloan, X. Wei, L. Zhan, and R. A. Irizarry. A benchmark for RNA-seq quantification pipelines. *Genome Biology*, 17:74, apr 2016. doi: 10.1186/s13059-016-0940-1. URL http://dx.doi.org/10.1186/s13059-016-0940-1.

L. Tian, X. Dong, S. Freytag, K.-A. Lê Cao, S. Su, A. JalalAbadi, D. Amann-Zalcenstein, T. S. Weber, A. Seidi, J. S. Jabbari, S. H. Naik, and M. E.

Ritchie. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nature Methods*, 16(6):479–487, may 2019. ISSN 1548-7091. doi: 10.1038/s41592-019-0425-8. URL http://www.nature.com/articles/s41592-019-0425-8.

Y. Toyooka, D. Shimosato, K. Murakami, K. Takahashi, and H. Niwa. Identification and characterization of subpopulations in undifferentiated ES cell culture. *Development*, 135(5):909–918, mar 2008. doi: 10.1242/dev.017400. URL http://dx.doi.org/10.1242/dev.017400.

C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515, may 2010. doi: 10.1038/nbt.1621. URL http://dx.doi.org/10.1038/nbt.1621.

P.-Y. Tung, J. D. Blischak, C. J. Hsiao, D. A. Knowles, J. E. Burnett, J. K. Pritchard, and Y. Gilad. Batch effects and the effective design of single-cell gene expression studies. *Scientific reports*, 7:39921, jan 2017. doi: 10.1038/srep39921. URL http://dx.doi.org/10.1038/srep39921.

E. A. Urban and R. J. Johnston. Buffering and amplifying transcriptional noise during cell fate specification. *Frontiers in genetics*, 9:591, nov 2018. ISSN 1664-8021. doi: 10.3389/fgene.2018.00591. URL https://www.frontiersin.org/article/10.3389/fgene.2018.00591/full.

S. Valadkhan, A. Mohammadi, C. Wachtel, and J. L. Manley. Protein-free spliceosomal snRNAs catalyze a reaction that resembles the first step of splicing. *RNA (New York)*, 13(12):2300–2311, dec 2007. doi: 10.1261/rna.626207. URL http://dx.doi.org/10.1261/rna.626207.

S. Valadkhan, A. Mohammadi, Y. Jaladat, and S. Geisler. Protein-free small nuclear RNAs catalyze a two-step splicing reaction. *Proceedings of the National Academy*

*of Sciences of the United States of America*, 106(29):11901–11906, jul 2009. doi: 10.1073/pnas.0902020106. URL http://dx.doi.org/10.1073/pnas.0902020106.

C. A. Vallejos, D. Risso, A. Scialdone, S. Dudoit, and J. C. Marioni. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nature Methods*, 14(6):565–571, jun 2017. ISSN 1548-7091. doi: 10.1038/nmeth.4292. URL http://www.nature.com/doifinder/10.1038/nmeth.4292.

D. van Dijk, R. Sharma, J. Nainys, K. Yim, P. Kathail, A. J. Carr, C. Burdziak, K. R. Moon, C. L. Chaffer, D. Pattabiraman, B. Bierie, L. Mazutis, G. Wolf, S. Krishnaswamy, and D. Pe'er. Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174(3):716–729.e27, jul 2018. ISSN 00928674. doi: 10.1016/j.cell.2018.05.061. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867418307244.

L. Velten, S. Anders, A. Pekowska, A. I. Järvelin, W. Huber, V. Pelechano, and L. M. Steinmetz. Single-cell polyadenylation site mapping reveals 3' isoform choice variability. *Molecular Systems Biology*, 11(6):812, jun 2015. doi: 10.15252/msb.20156198. URL http://dx.doi.org/10.15252/msb.20156198.

J. Verne. Journey to the centre of the earth. 1864.

S. Vickovic, G. Eraslan, J. Klughammer, L. Stenbeck, F. Salmen, T. Aijo, R. Bonneau, L. Bergenstraahle, J. Gould, M. Ronaghi, J. Frisen, J. Lundeberg, A. Regev, and P. L. Staahl. High-density spatial transcriptomics arrays for in situ tissue profiling. *BioRxiv*, feb 2019. doi: 10.1101/563338. URL http://biorxiv.org/lookup/doi/10.1101/563338.

F. Wagner, Y. Yan, and I. Yanai. K-nearest neighbor smoothing for high-throughput single-cell RNA-seq data. *BioRxiv*, nov 2017. doi: 10.1101/217737. URL http://biorxiv.org/lookup/doi/10.1101/217737.

Z. Waks, A. M. Klein, and P. A. Silver. Cell-to-cell variability of alternative RNA splicing. *Molecular Systems Biology*, 7:506, jul 2011. doi: 10.1038/msb.2011.32. URL http://dx.doi.org/10.1038/msb.2011.32.

R. Wan, C. Yan, R. Bai, G. Huang, and Y. Shi. Structure of a yeast catalytic step i spliceosome at 3.4 å resolution. *Science*, 353(6302):895–904, aug 2016a. doi: 10.1126/science.aag2235. URL http://dx.doi.org/10.1126/science.aag2235.

R. Wan, C. Yan, R. Bai, L. Wang, M. Huang, C. C. L. Wong, and Y. Shi. The 3.8 å structure of the u4/u6.u5 tri-snRNP: Insights into spliceosome assembly and catalysis. *Science*, 351(6272):466–475, jan 2016b. doi: 10.1126/science.aad6466. URL http://dx.doi.org/10.1126/science.aad6466.

E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, nov 2008. ISSN 1476-4687. doi: 10.1038/nature07509. URL http://dx.doi.org/10.1038/nature07509.

L. Wang, S. Wang, and W. Li. RSeQC: quality control of RNA-seq experiments. *Bioinformatics*, 28(16):2184–2185, aug 2012. doi: 10.1093/bioinformatics/bts356. URL http://dx.doi.org/10.1093/bioinformatics/bts356.

D. A. Wassarman and J. A. Steitz. Interactions of small nuclear RNA's with precursor messenger RNA during in vitro splicing. *Science*, 257(5078):1918–1925, sep 1992. doi: 10.1126/science.1411506. URL http://dx.doi.org/10.1126/science.1411506.

A. P. M. Weber. Discovering new biology through sequencing of RNA. *Plant Physiology*, 169(3):1524–1531, nov 2015. doi: 10.1104/pp.15.01081. URL http://dx.doi.org/10.1104/pp.15.01081.

G. Weber, S. Trowitzsch, B. Kastner, R. Lührmann, and M. C. Wahl. Functional organization of the sm core in the crystal structure of human u1 snRNP. *The EMBO Journal*, 29(24):4172–4184, dec 2010. doi: 10.1038/emboj.2010.295. URL http://dx.doi.org/10.1038/emboj.2010.295.

J. Westoby, M. S. Herrera, A. C. Ferguson-Smith, and M. Hemberg. Simulation-based benchmarking of isoform quantification in single-cell RNA-seq. *Genome Biology*, 19(1):191, nov 2018a. ISSN 1474-760X. doi: 10.

1186/s13059-018-1571-5. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1571-5.

J. Westoby, M. Sjoberg, A. Ferguson-Smith, and M. Hemberg. Simulation based benchmarking of isoform quantification in single-cell RNA-seq. *BioRxiv*, jan 2018b. doi: 10.1101/248716. URL http://biorxiv.org/lookup/doi/10.1101/248716.

J. Westoby, P. Artemov, M. Hemberg, and A. C. Ferguson-Smith. Obstacles to studying alternative splicing using scRNA-seq. *BioRxiv*, oct 2019. doi: 10.1101/797951. URL http://biorxiv.org/lookup/doi/10.1101/797951.

S. Wiesner, G. Stier, M. Sattler, and M. J. Macias. Solution structure and ligand recognition of the WW domain pair of the yeast splicing factor prp40. *Journal of Molecular Biology*, 324(4):807–822, dec 2002. doi: 10.1016/s0022-2836(02)01145-2. URL http://dx.doi.org/10.1016/s0022-2836(02)01145-2.

O. Wilde. The importance of being earnest. 1895.

E. Wolf, B. Kastner, J. Deckert, C. Merz, H. Stark, and R. Lührmann. Exon, intron and splice site locations in the spliceosomal b complex. *The EMBO Journal*, 28 (15):2283–2292, aug 2009. doi: 10.1038/emboj.2009.171. URL http://dx.doi.org/10.1038/emboj.2009.171.

S. L. Wolock, R. Lopez, and A. M. Klein. Scrublet: Computational identification of cell doublets in single-cell transcriptomic data. *Cell Systems*, 8(4):281–291.e9, apr 2019. doi: 10.1016/j.cels.2018.11.005. URL http://dx.doi.org/10.1016/j.cels.2018.11.005.

A. R. Wu, N. F. Neff, T. Kalisky, P. Dalerba, B. Treutlein, M. E. Rothenberg, F. M. Mburu, G. L. Mantalas, S. Sim, M. F. Clarke, and S. R. Quake. Quantitative assessment of single-cell RNA-sequencing methods. *Nature Methods*, 11(1):41–46, jan 2014. doi: 10.1038/nmeth.2694. URL http://dx.doi.org/10.1038/nmeth.2694.

J. R. Wyatt, E. J. Sontheimer, and J. A. Steitz. Site-specific cross-linking of mammalian u5 snRNP to the 5' splice site before the first step of pre-mRNA splicing. *Genes & Development*, 6(12B):2542–2553, dec 1992. doi: 10.1101/gad.6.12b.2542. URL http://dx.doi.org/10.1101/gad.6.12b.2542.

T. Yamazaki, L. Liu, D. Lazarev, A. Al-Zain, V. Fomin, P. L. Yeung, S. M. Chambers, C.-W. Lu, L. Studer, and J. L. Manley. TCF3 alternative splicing controlled by hnRNP h/f regulates e-cadherin expression and hESC pluripotency. *Genes & Development*, 32(17-18):1161–1174, sep 2018. doi: 10.1101/gad.316984.118. URL http://dx.doi.org/10.1101/gad.316984.118.

C. Yan, J. Hang, R. Wan, M. Huang, C. C. L. Wong, and Y. Shi. Structure of a yeast spliceosome at 3.6-angstrom resolution. *Science*, 349(6253):1182–1191, sep 2015. ISSN 0036-8075. doi: 10.1126/science.aac7629. URL http://www.sciencemag.org/cgi/doi/10.1126/science.aac7629.

C. Yan, R. Wan, R. Bai, G. Huang, and Y. Shi. Structure of a yeast activated spliceosome at 3.5 å resolution. *Science*, 353(6302):904–911, aug 2016. doi: 10.1126/science.aag0291. URL http://dx.doi.org/10.1126/science.aag0291.

C. Yan, R. Wan, R. Bai, G. Huang, and Y. Shi. Structure of a yeast step II catalytically activated spliceosome. *Science*, 355(6321):149–155, jan 2017. doi: 10.1126/science.aak9979. URL http://dx.doi.org/10.1126/science.aak9979.

H. Yoshida, S.-Y. Park, T. Oda, T. Akiyoshi, M. Sato, M. Shirouzu, K. Tsuda, K. Kuwasako, S. Unzai, Y. Muto, T. Urano, and E. Obayashi. A novel 3' splice site recognition by the two zinc fingers in the U2AF small subunit. *Genes & Development*, 29(15):1649–1660, aug 2015. doi: 10.1101/gad.267104.115. URL http://dx.doi.org/10.1101/gad.267104.115.

L. Zappia, B. Phipson, and A. Oshlack. Splatter: Simulation of single-cell RNA sequencing data. *BioRxiv*, may 2017a. doi: 10.1101/133173. URL http://biorxiv.org/lookup/doi/10.1101/133173.

L. Zappia, B. Phipson, and A. Oshlack. Splatter: simulation of single-cell RNA sequencing data. *Genome Biology*, 18(1):174, sep 2017b. doi: 10.1186/s13059-017-1305-0. URL http://dx.doi.org/10.1186/s13059-017-1305-0.

L. Zappia, B. Phipson, and A. Oshlack. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Computational Biology*, 14(6): e1006245, jun 2018. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006245. URL http://dx.plos.org/10.1371/journal.pcbi.1006245.

C. Zhang, B. Zhang, L.-L. Lin, and S. Zhao. Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics*, 18(1):583, aug 2017. doi: 10.1186/s12864-017-4002-1. URL http://dx.doi.org/10.1186/s12864-017-4002-1.

X. Zhang, C. Yan, X. Zhan, L. Li, J. Lei, and Y. Shi. Structure of the human activated spliceosome in three conformational states. *Cell Research*, 28(3):307–322, mar 2018. doi: 10.1038/cr.2018.14. URL http://dx.doi.org/10.1038/cr.2018.14.

X. Zhang, X. Zhan, C. Yan, W. Zhang, D. Liu, J. Lei, and Y. Shi. Structures of the human spliceosomes before and after release of the ligated exon. *Cell Research*, 29(4):274–285, apr 2019. doi: 10.1038/s41422-019-0143-x. URL http://dx.doi.org/10.1038/s41422-019-0143-x.

Z. Zhao, J. Tu, Z. Lu, and S. Liu. Dominant isoform in alternative splicing in HeLa s3 cell line revealed by single-cell RNA-seq. In *Proceedings of the 7th International Conference on Computational Systems-Biology and Bioinformatics - CSBio '16*, pages 1–7, New York, New York, USA, dec 2016. ACM Press. ISBN 9781450347945. doi: 10.1145/3029375.3029376. URL http://dl.acm.org/citation.cfm?doid=3029375.3029376.

G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W.

Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, and J. H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:14049, jan 2017. doi: 10.1038/ncomms14049. URL http://dx.doi.org/10.1038/ncomms14049.

L. Zhou, J. Hang, Y. Zhou, R. Wan, G. Lu, P. Yin, C. Yan, and Y. Shi. Crystal structures of the lsm complex bound to the 3' end sequence of u6 small nuclear RNA. *Nature*, 506(7486):116–120, feb 2014. doi: 10.1038/nature12803. URL http://dx.doi.org/10.1038/nature12803.

C. Ziegenhain, B. Vieth, S. Parekh, B. Reinius, A. Guillaumet-Adkins, M. Smets, H. Leonhardt, H. Heyn, I. Hellmann, and W. Enard. Comparative analysis of single-cell RNA sequencing methods. *Molecular Cell*, 65(4):631–643.e4, feb 2017. doi: 10.1016/j.molcel.2017.01.023. URL http://dx.doi.org/10.1016/j.molcel.2017.01.023.

# 7

# Appendix 1

This appendix contains quality control and other descriptive plots from chapter 2.

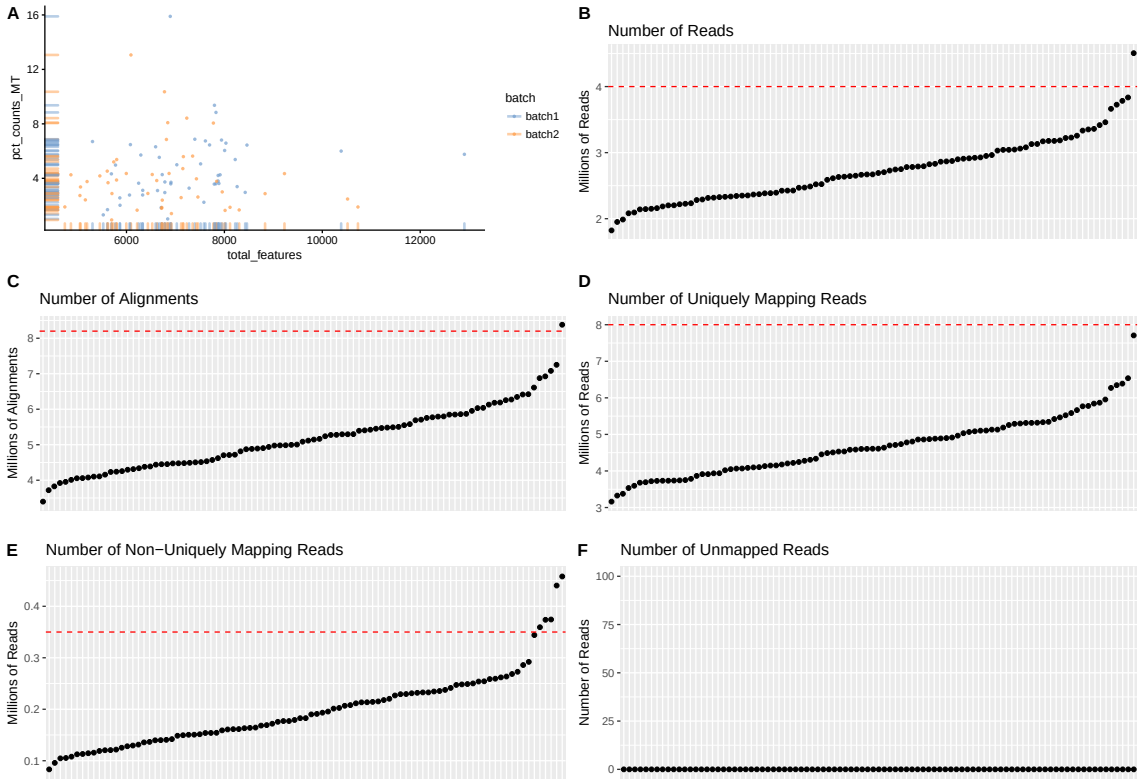


Figure 7.1: Plots of quality control statistics for the BLUEPRINT B lymphocytes. In all of these plots, one point represents one cell. Based on these plots, cells with more than 10% of reads mapping to mitochondrial RNA, more than 4 million reads, more than 8.2 million alignments, more than 8 million uniquely mapping reads or more than 350,000 non-uniquely mapping reads were removed. Dashed red lines indicate the thresholds selected to remove cells. **A** Percentage of reads mapping to mitochondrial RNA. Graph produced using the scater package(McCarthy et al., 2017). **B** Number of reads per cell. **C** Number of alignments per cell. **D** Number of uniquely mapping reads per cell. **E** Number of non-uniquely mapping reads per cell. **F** Number of unmapped reads.

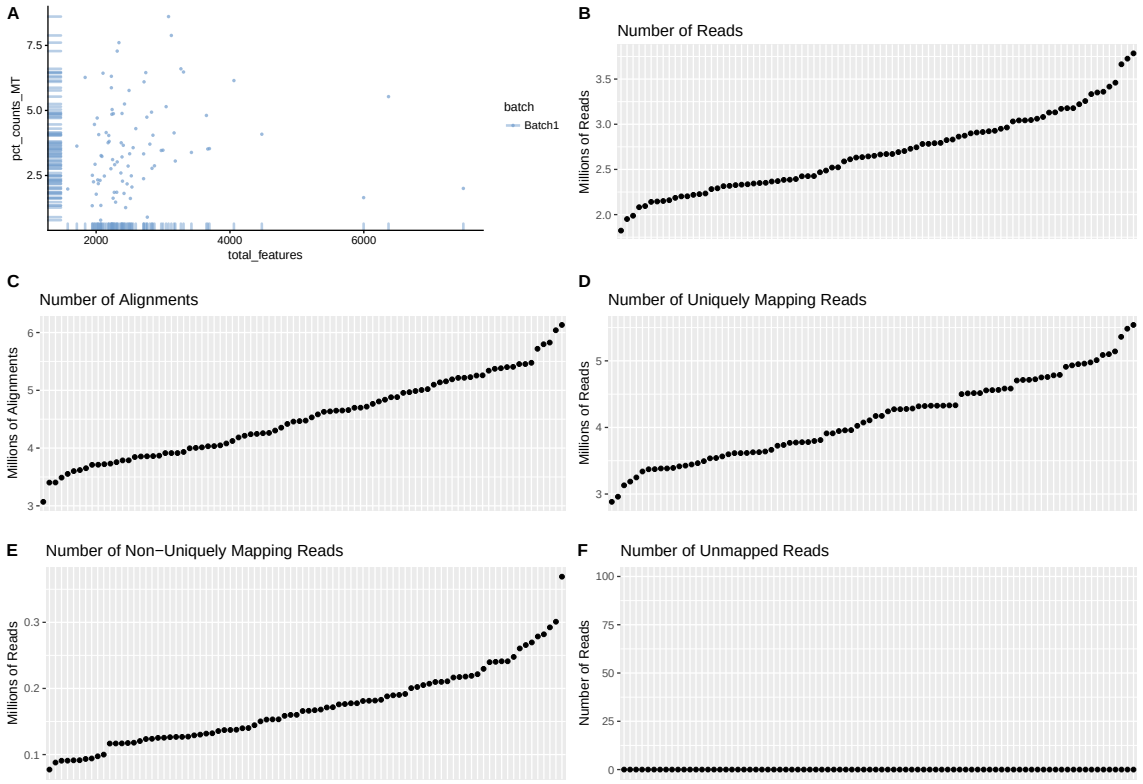


Figure 7.2: Plots of quality control statistics for the RSEM(Li and Dewey, 2011) simulated data based on the BLUEPRINT B lymphocytes. In all of these plots, one point represents one cell. Based on these plots, cells with more than 10% of reads mapping to mitochondrial RNA were removed. **A** Percentage of reads mapping to mitochondrial RNA. Graph produced using the scater package. **B** Number of reads per cell. **C** Number of alignments per cell. **D** Number of uniquely mapping reads per cell. **E** Number of non-uniquely mapping reads per cell. **F** Number of unmapped reads.

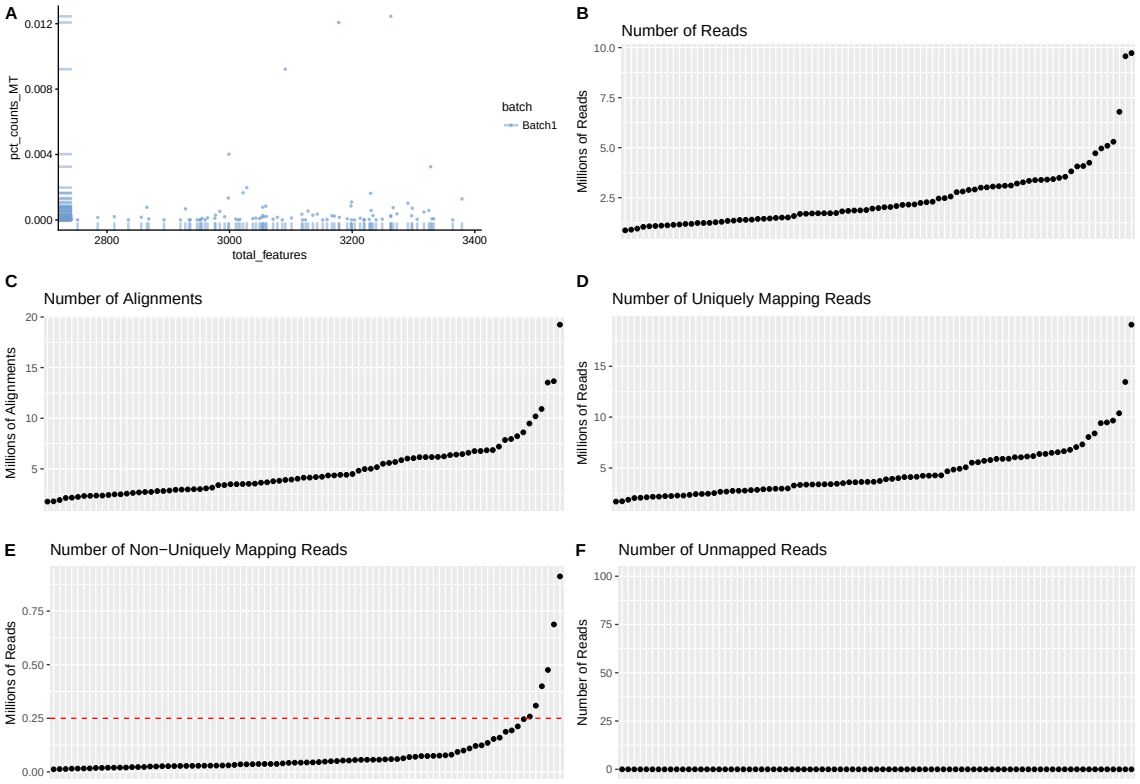


Figure 7.3: Plots of quality control statistics for the Splatter(Zappia et al., 2017b) and Polyester(Frazee et al., 2015) 3' bias simulated data based on the BLUEPRINT B lymphocytes. In all of these plots, one point represents one cell. Based on these plots, cells with more than 250,000 non-uniquely mapping reads were removed. Dashed red lines indicate the thresholds selected to remove cells. **A** Percentage of reads mapping to mitochondrial RNA. Graph produced using the scater package. **B** Number of reads per cell. **C** Number of alignments per cell. **D** Number of uniquely mapping reads per cell. **E** Number of non-uniquely mapping reads per cell. **F** Number of unmapped reads.

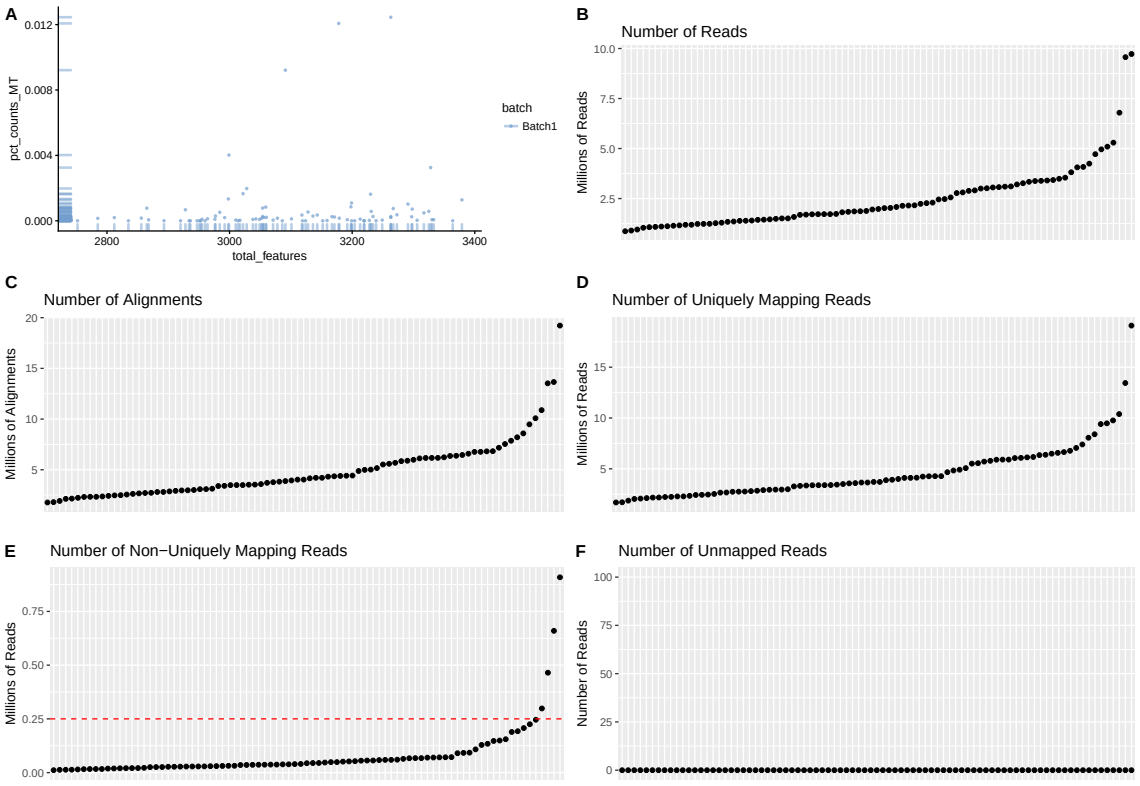


Figure 7.4: Plots of quality control statistics for the Splatter and Polyester simulated data based on the BLUEPRINT B lymphocytes, simulated with no coverage bias. In all of these plots, one point represents one cell. Based on these plots, no poor quality cells were removed. Based on these plots, cells with more than 250,000 non uniquely mapping reads were removed. Dashed red lines indicate the thresholds selected to remove cells. **A** Percentage of reads mapping to mitochondrial RNA. Graph produced using the scater package. **B** Number of reads per cell. **C** Number of alignments per cell. **D** Number of uniquely mapping reads per cell. **E** Number of non-uniquely mapping reads per cell. **F** Number of unmapped reads.

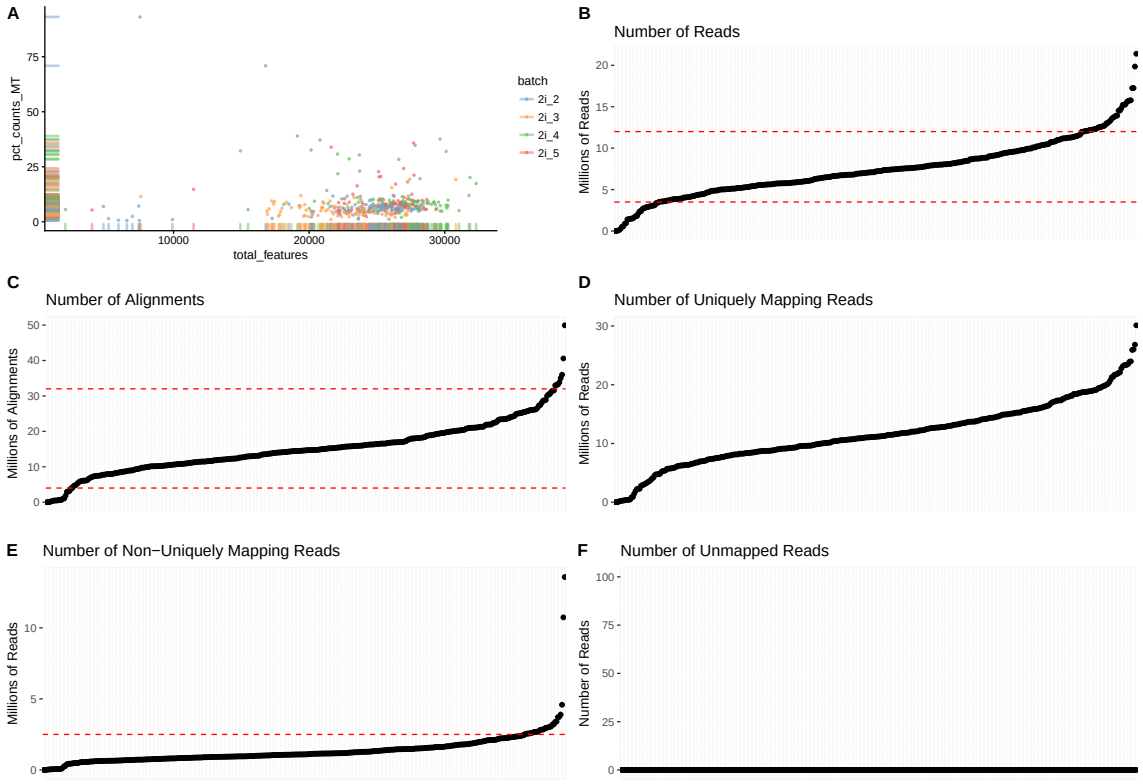


Figure 7.5: Plots of quality control statistics for mESCs grown in standard 2i media + LIF, published by Kolodziejczyk et al.(Kolodziejczyk et al., 2015) In all of these plots, one point represents one cell. Based on these plots, cells with more than 10% of reads mapping to mitochondrial RNA, more than 12 million or less than 3.5 million reads, more than 32 million or less than 4 million alignments, or more than 2.5 million non-uniquely mapping reads were removed. Dashed red lines indicate the thresholds selected to remove cells. **A** Percentage of reads mapping to mitochondrial RNA. Graph produced using the scater package. **B** Number of reads per cell. **C** Number of alignments per cell. **D** Number of uniquely mapping reads per cell. **E** Number of non-uniquely mapping reads per cell. F: Number of unmapped reads.

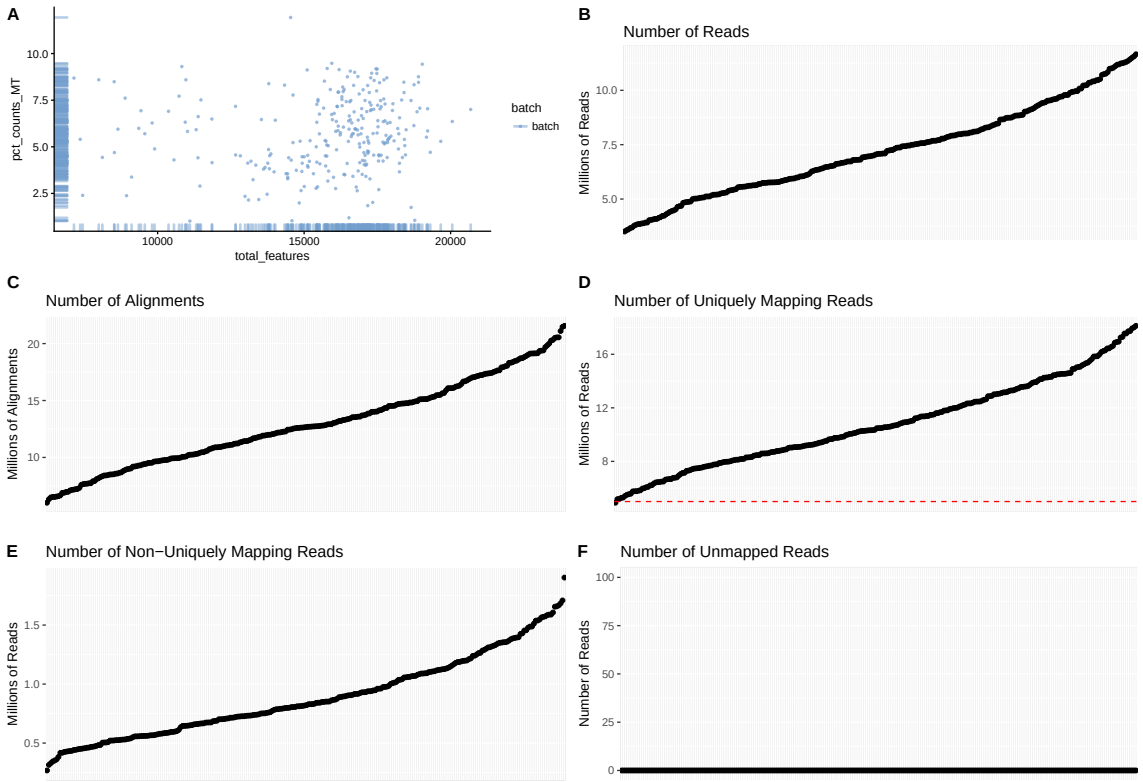


Figure 7.6: Plots of quality control statistics for simulated mESCs. In all of these plots, one point represents one cell. Based on these plots, cells with more than 10% of reads mapping to mitochondrial RNA or less than 5 million uniquely mapping reads were removed. Dashed red lines indicate the thresholds selected to remove cells. **A** Percentage of reads mapping to mitochondrial RNA. Graph produced using the scater package. **B** Number of reads per cell. **C** Number of alignments per cell. **D** Number of uniquely mapping reads per cell. **E** Number of non-uniquely mapping reads per cell. **F** Number of unmapped reads.

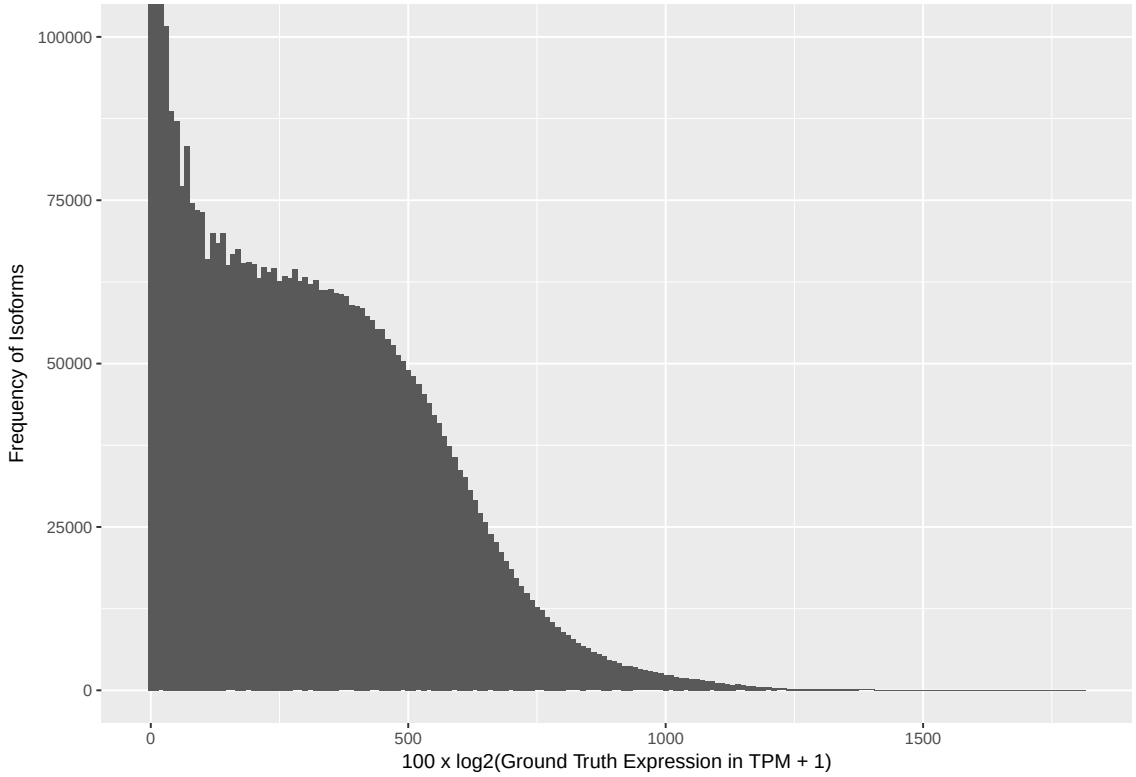


Figure 7.7: Histogram of ground truth expression values for simulated mESCs.

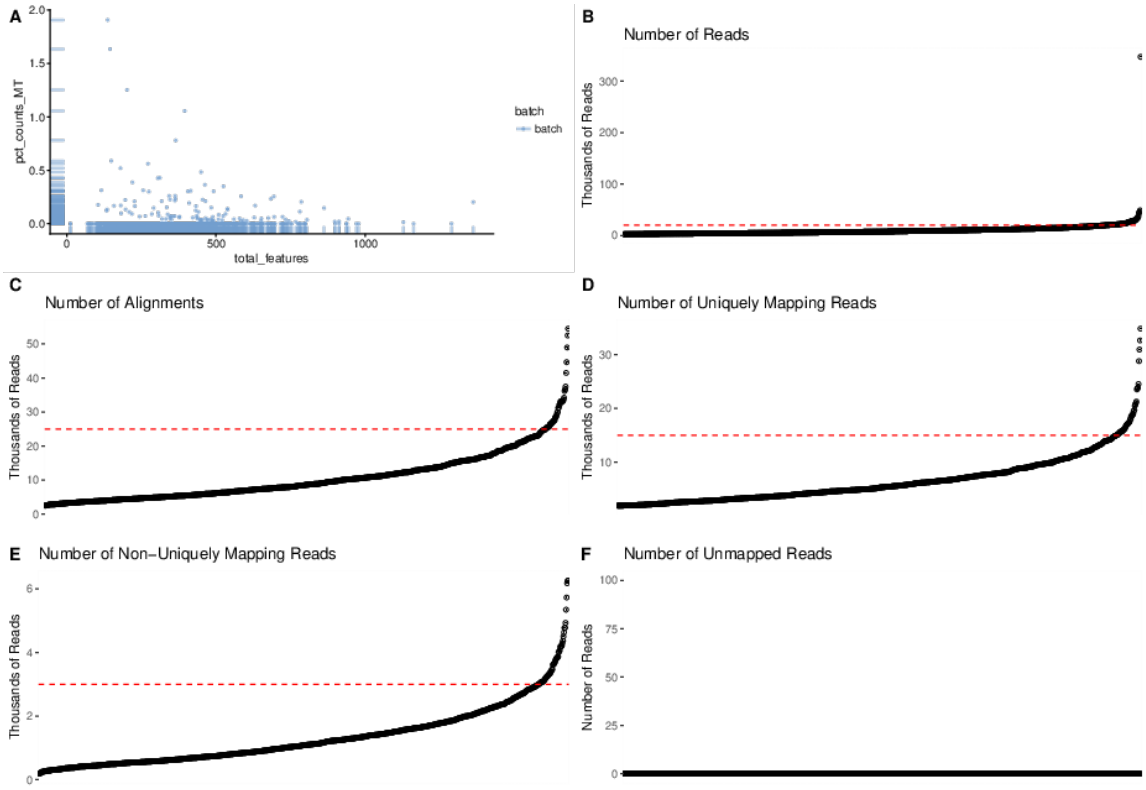


Figure 7.8: Plots of quality control statistics for 1000 randomly selected Drop-seq cells. In all of these plots, one point represents one cell. Based on these plots, cells with more than 10% of reads mapping to mitochondrial RNA, more than 20,000 reads, more than 25,000 alignments, more than 15,000 uniquely mapping reads or more than 3,000 non-uniquely mapping reads were removed. Dashed red lines indicate the thresholds selected to remove cells. **A** Percentage of reads mapping to mitochondrial RNA. Graph produced using the scater package. **B** Number of reads per cell. **C** Number of alignments per cell. **D** Number of uniquely mapping reads per cell. **E** Number of non-uniquely mapping reads per cell. **F** Number of unmapped reads.

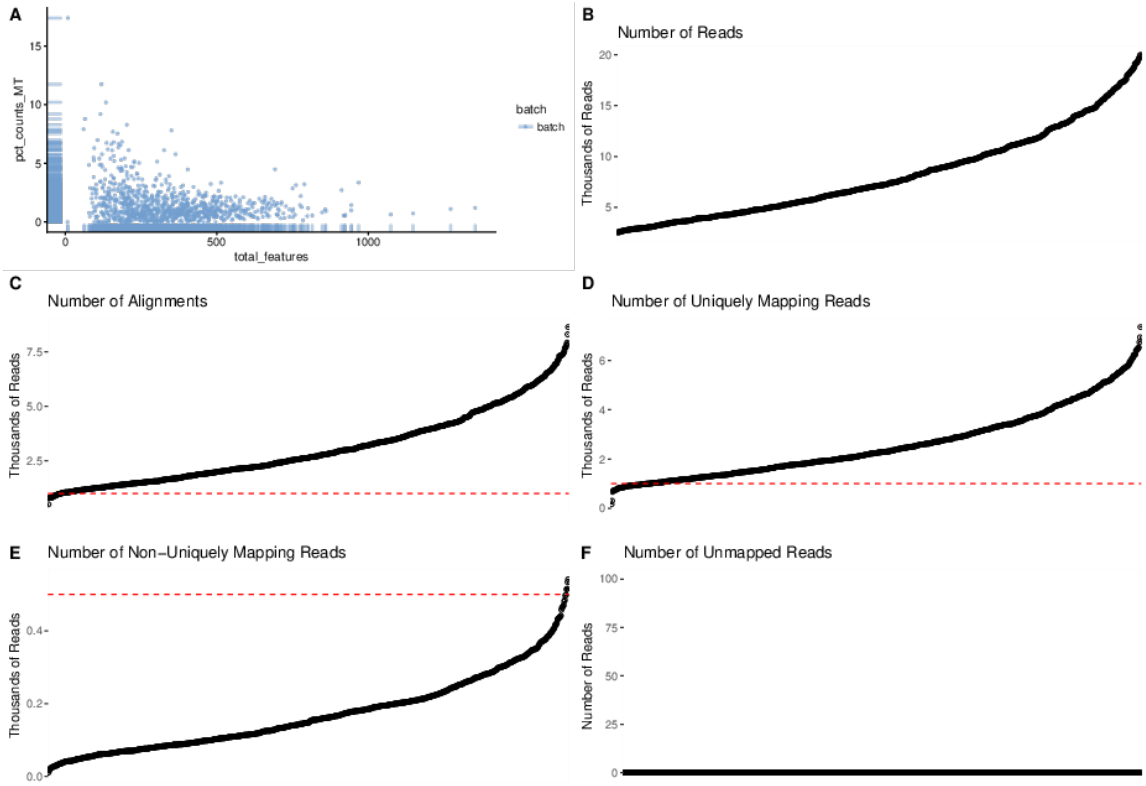


Figure 7.9: Plots of quality control statistics for the simulated Drop-seq cells. In all of these plots, one point represents one cell. Based on these plots, cells with more than 10% of reads mapping to mitochondrial RNA, less than 1,000 alignments, less than 1,000 uniquely mapping reads or more than 500 non-uniquely mapping reads were removed. Dashed red lines indicate the thresholds selected to remove cells. The number of alignments differs substantially in the simulated data compared to the real data. One explanation for this could be inaccurate quantification of the number of alignments. **A** Percentage of reads mapping to mitochondrial RNA. Graph produced using the scater package. **B** Number of reads per cell. **C** Number of alignments per cell. **D** Number of uniquely mapping reads per cell. **E** Number of non-uniquely mapping reads per cell. **F** Number of unmapped reads.

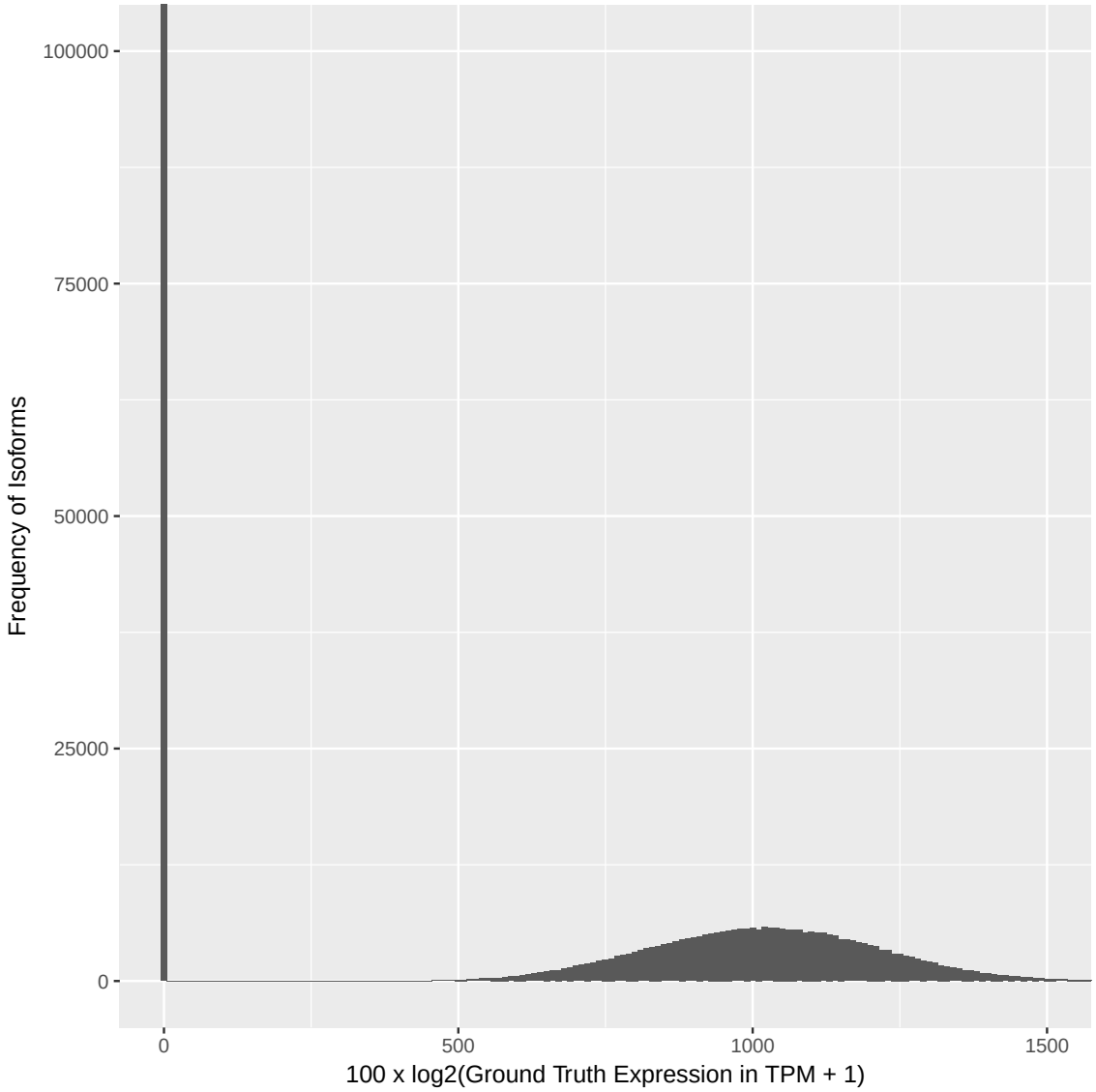


Figure 7.10: Histogram of ground truth expression values for simulated Drop-seq cells.

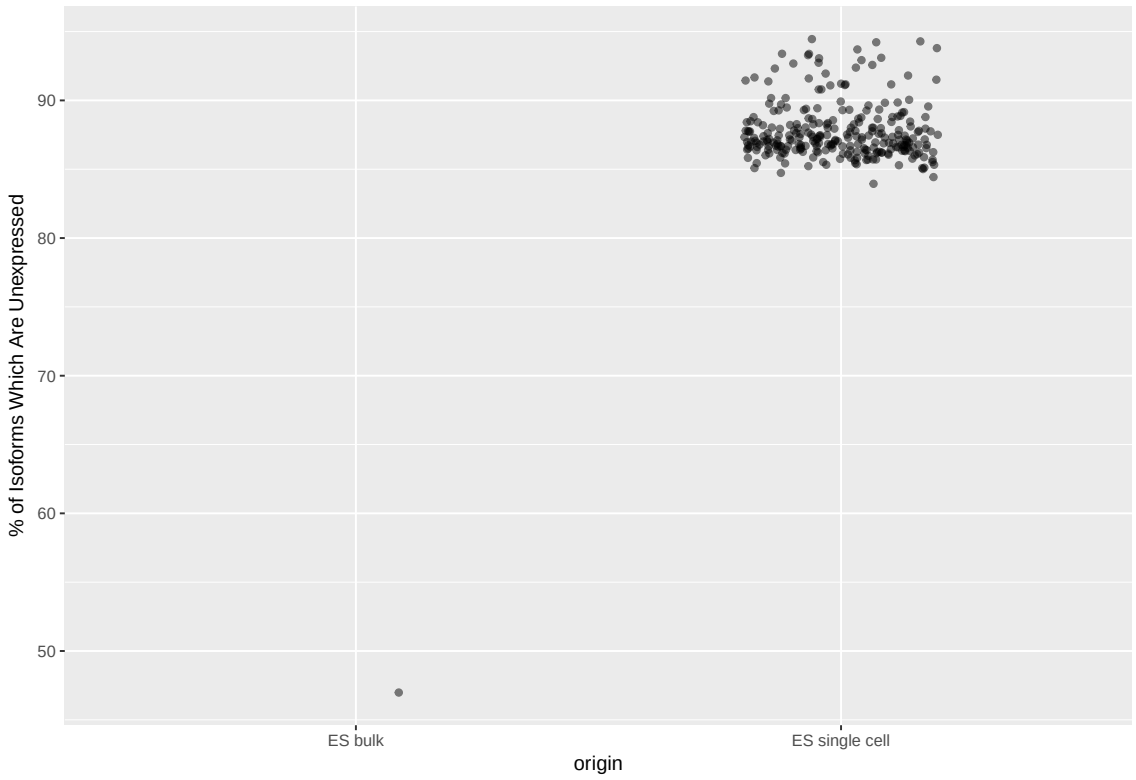


Figure 7.11: Comparison of the percentage of isoforms which are unexpressed (ie. have zero expression) in Kolodziejczyk et al. ES cell bulk and scRNA-seq data. For the single cell data, each point represents a simulated single cell. For the bulk data, each point represents a single simulated bulk RNA-seq sample.

# 8

# Appendix 2

This appendix contains simulation results from chapter 3 for pluripotency factors not predicted to express differing numbers of isoforms under different culture conditions.

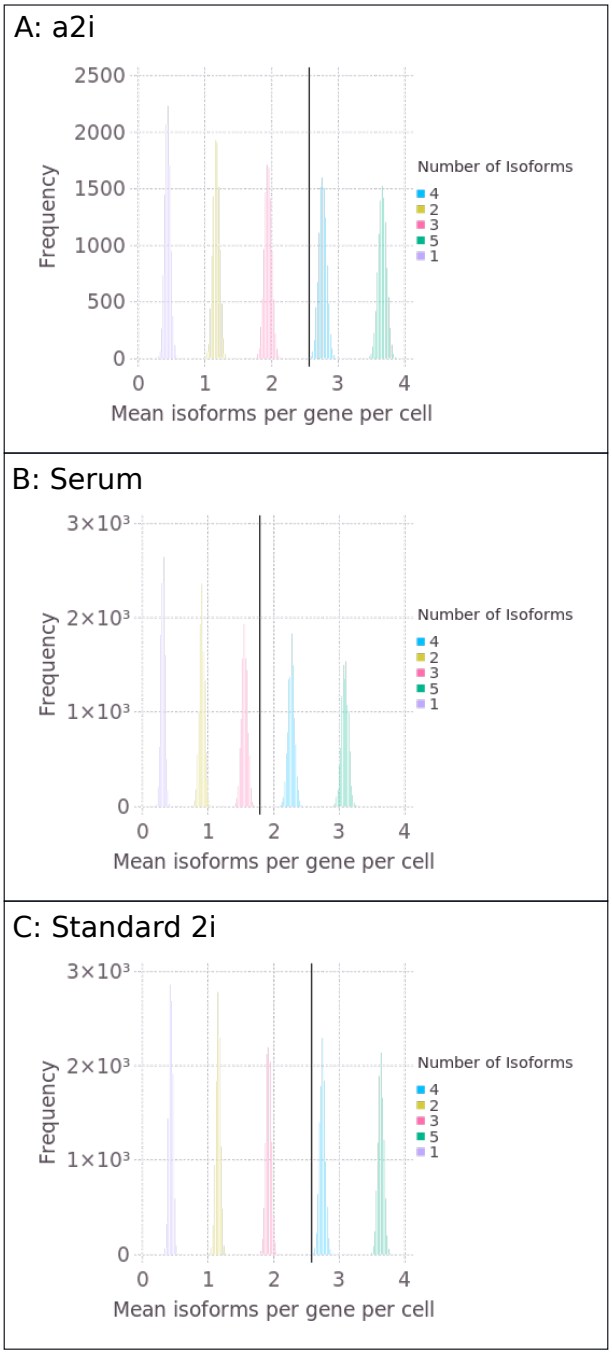


Figure 8.1: Simulation results for Esrrb gene in mESCs cultured in **A** a2i culture conditions, **B** serum and **C** standard 2i culture conditions. The vertical black line on each plot represents the mean number of isoforms detected per cell in the real data.

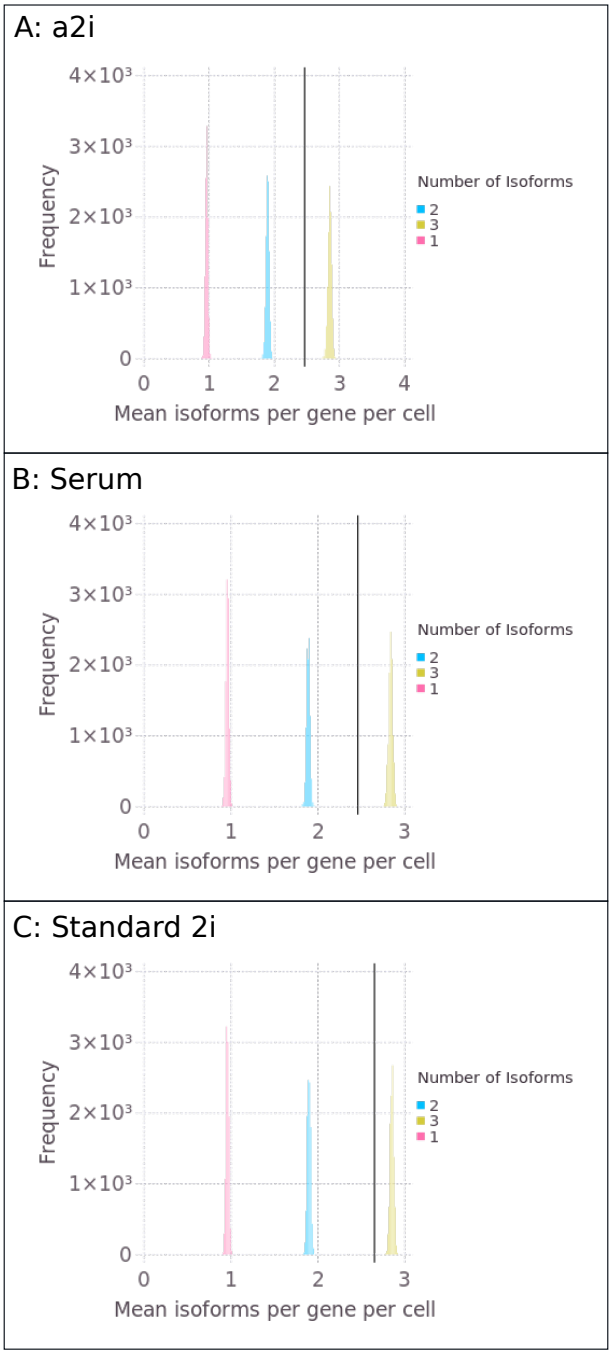


Figure 8.2: Simulation results for Nanog gene in mESCs cultured in **A** a2i culture conditions, **B** serum and **C** standard 2i culture conditions. The vertical black line on each plot represents the mean number of isoforms detected per cell in the real data.

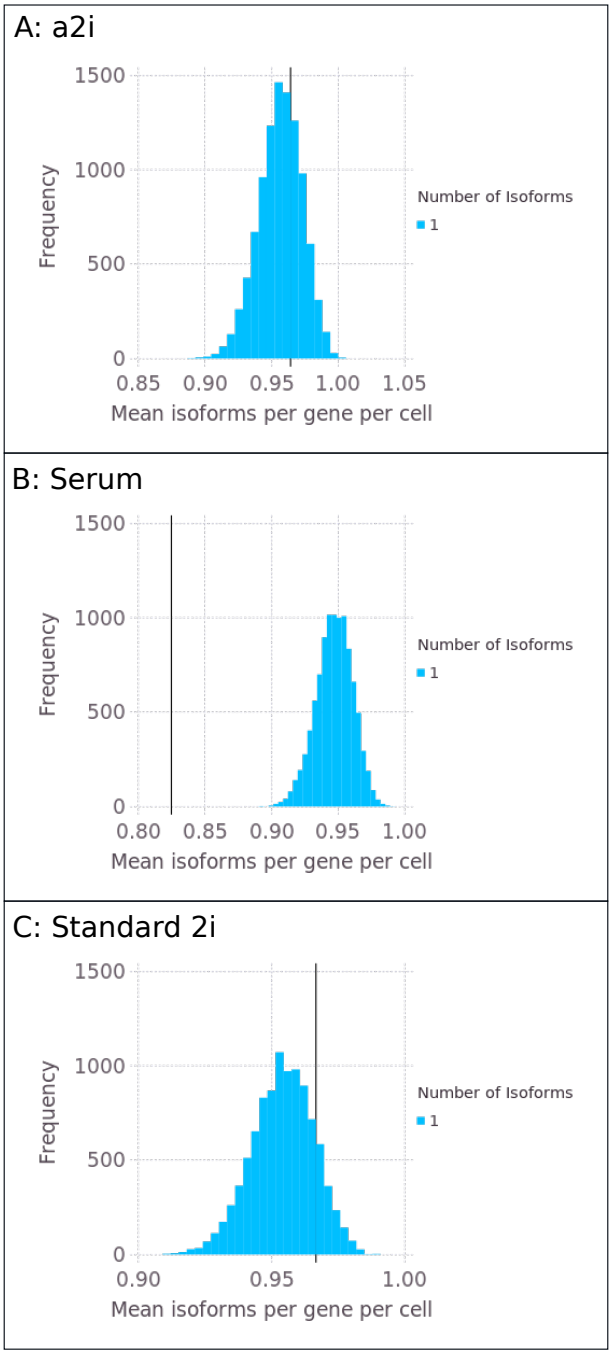


Figure 8.3: Simulation results for Nr0b1 gene in mESCs cultured in **A** a2i culture conditions, **B** serum and **C** standard 2i culture conditions. The vertical black line on each plot represents the mean number of isoforms detected per cell in the real data.

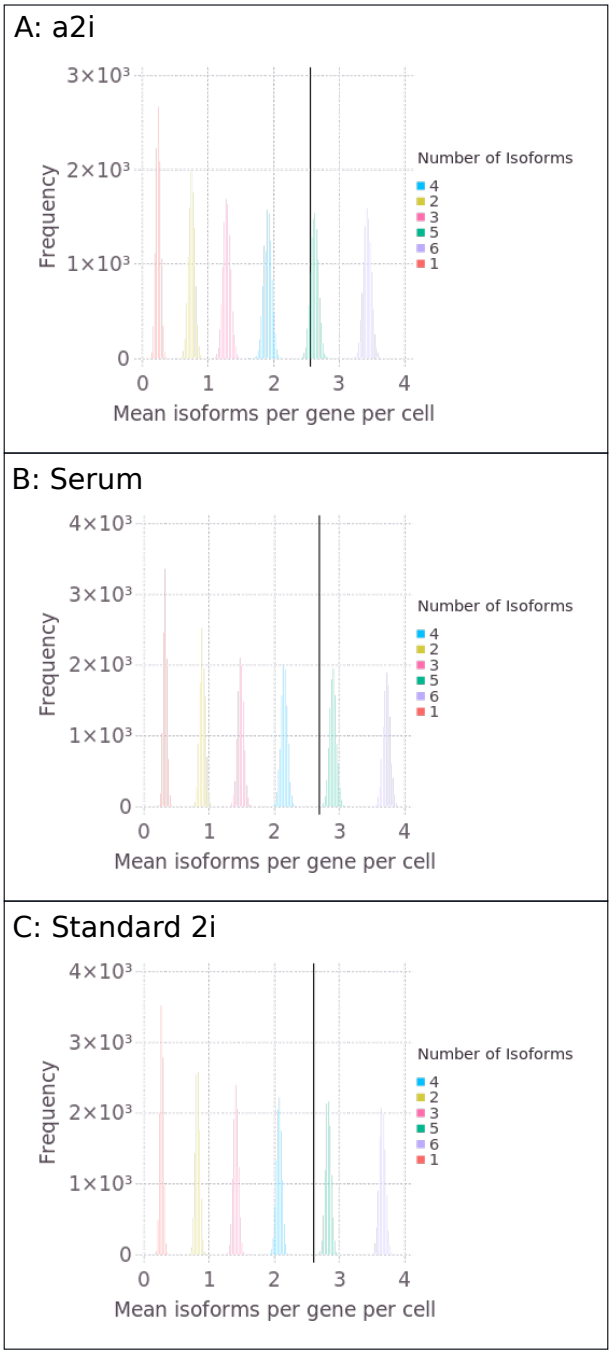


Figure 8.4: Simulation results for Sall4 gene in mESCs cultured in **A** a2i culture conditions, **B** serum and **C** standard 2i culture conditions. The vertical black line on each plot represents the mean number of isoforms detected per cell in the real data.

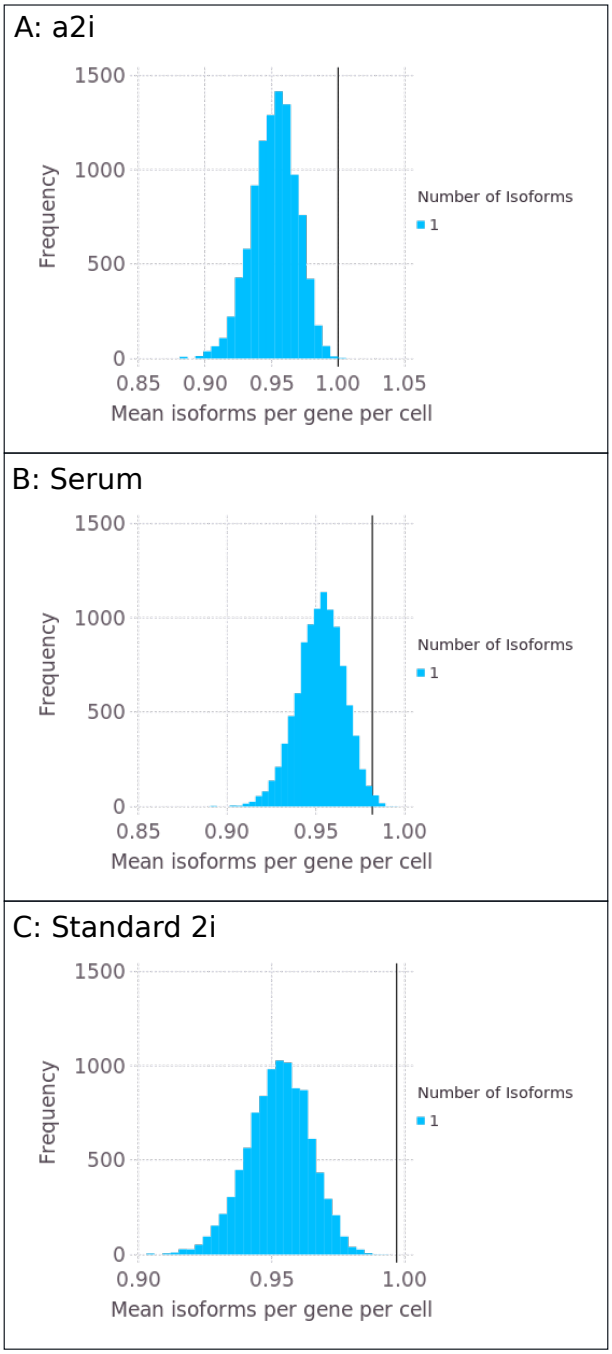


Figure 8.5: Simulation results for Sox2 gene in mESCs cultured in **A** a2i culture conditions, **B** serum and **C** standard 2i culture conditions. The vertical black line on each plot represents the mean number of isoforms detected per cell in the real data.

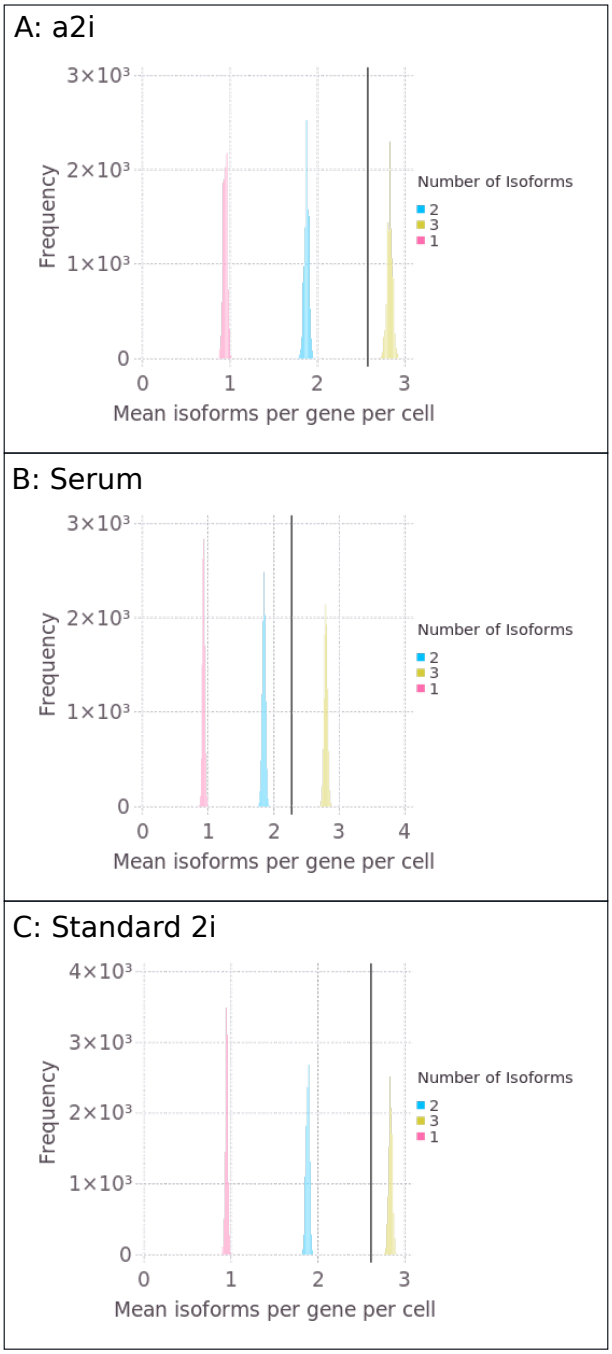


Figure 8.6: Simulation results for Zfp42 gene in mESCs cultured in **A** a2i culture conditions, **B** serum and **C** standard 2i culture conditions. The vertical black line on each plot represents the mean number of isoforms detected per cell in the real data.

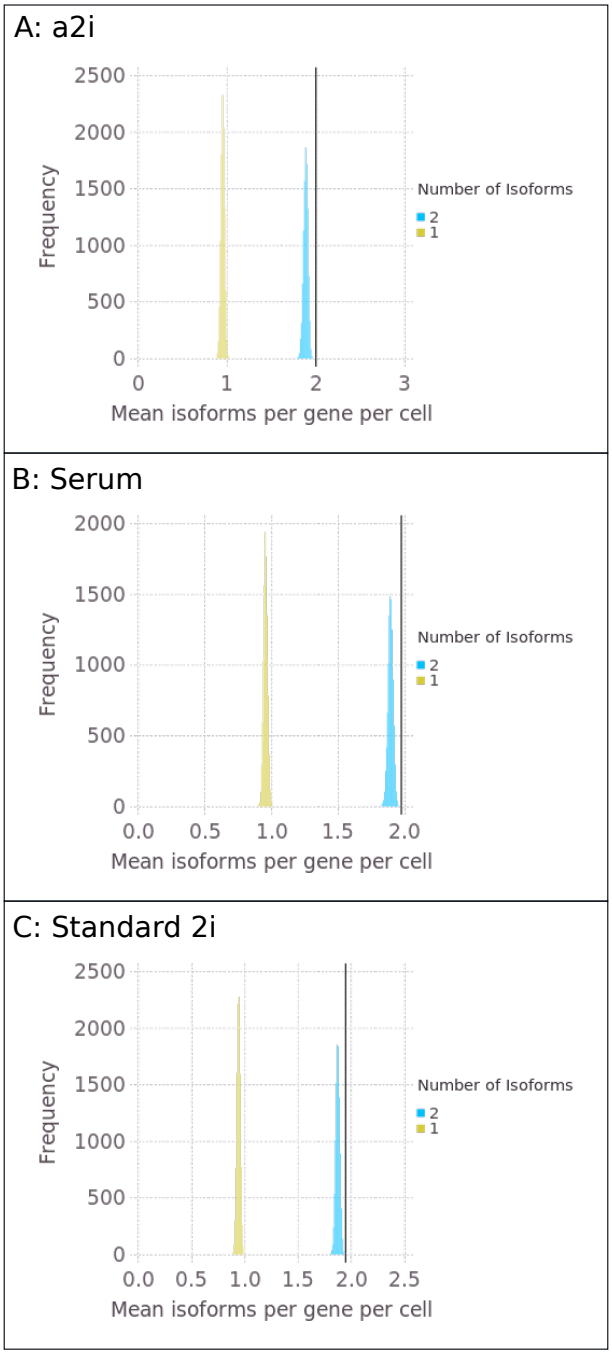


Figure 8.7: Simulation results for Zfp281 gene in mESCs cultured in **A** a2i culture conditions, **B** serum and **C** standard 2i culture conditions. The vertical black line on each plot represents the mean number of isoforms detected per cell in the real data.

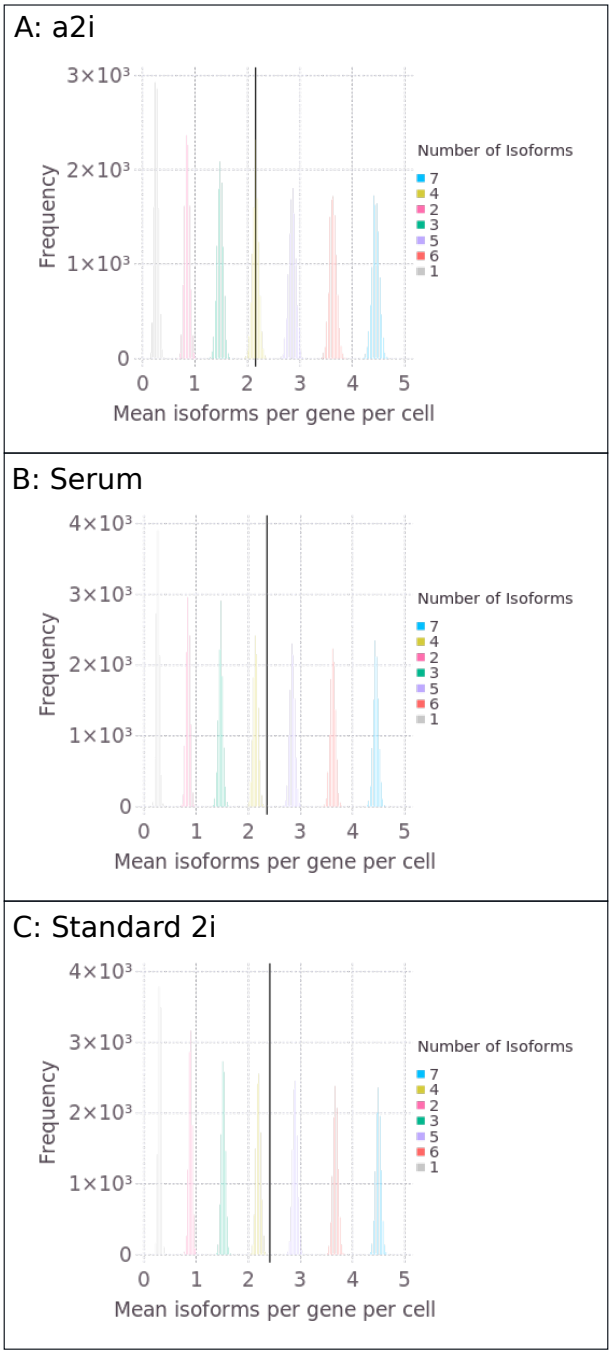


Figure 8.8: Simulation results for Zfx gene in mESCs cultured in **A** a2i culture conditions, **B** serum and **C** standard 2i culture conditions. The vertical black line on each plot represents the mean number of isoforms detected per cell in the real data.

# 9

# Appendix 3

Tables of statistical results and additional results figures from chapter 4 are presented in this chapter.

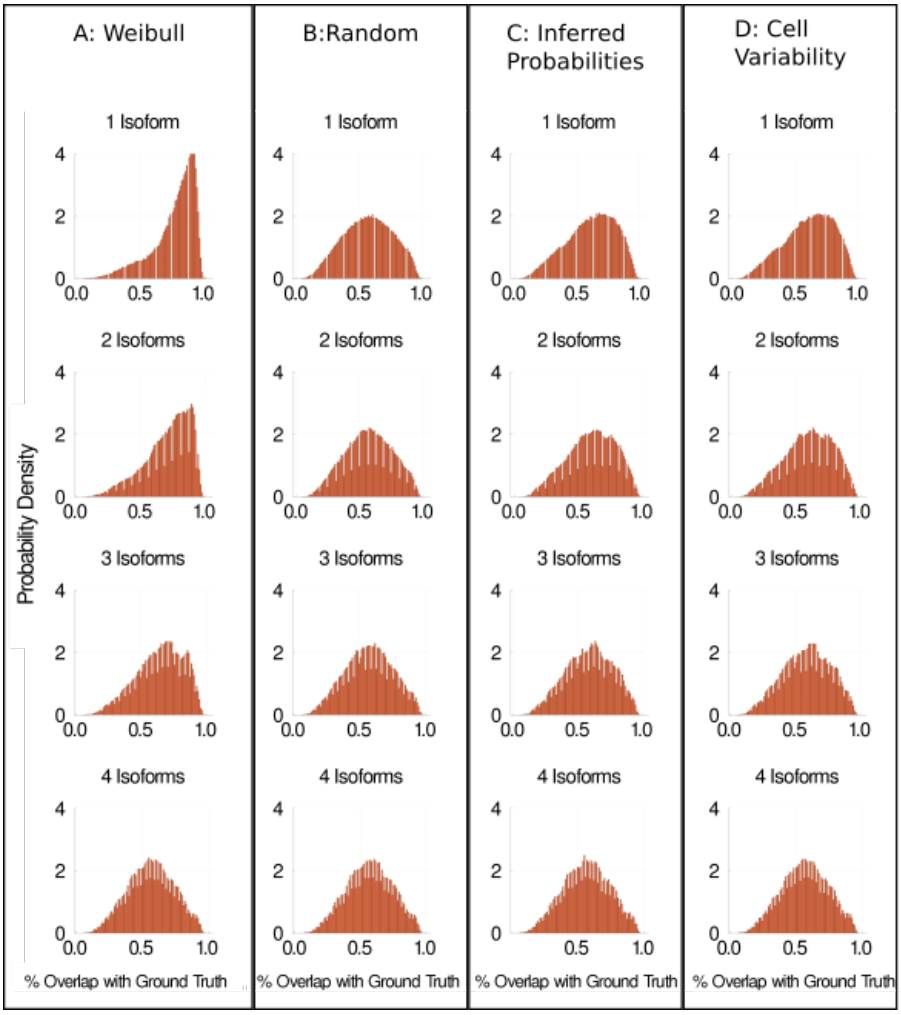


Figure 9.1: Distributions of the overlap fraction with the ground truth when the **A** Weibull model (Bacher et al., 2017; Hu et al., 2017), **B** random model, **C** inferred probabilities model and **D** cell variability model of isoform choice is used. All distributions are for H1 cells sequenced at approximately 4 million reads per cell. See the main text for a detailed description of each model.

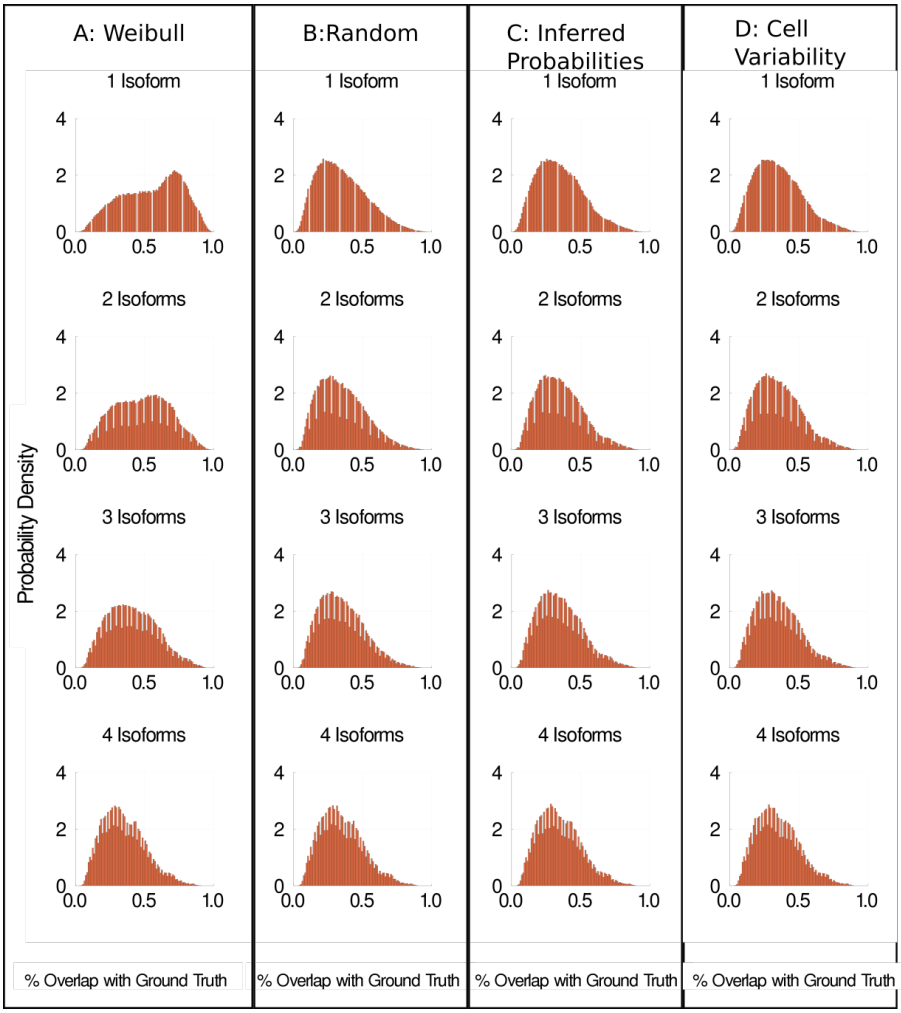


Figure 9.2: Different models of isoform choice alter our ability to detect isoforms. **A** Distributions of overlap fraction with the ground truth for H1 hESCs sequenced at approximately 1 million reads per cell using the Weibull model of isoform choice (Bacher et al., 2017; Hu et al., 2017). **B** shows the same distributions when the random model is used. **C** shows the distributions when the inferred probabilities model is used. **D** shows the distributions when the cell variability model is used. See the main text for a detailed description of each model.

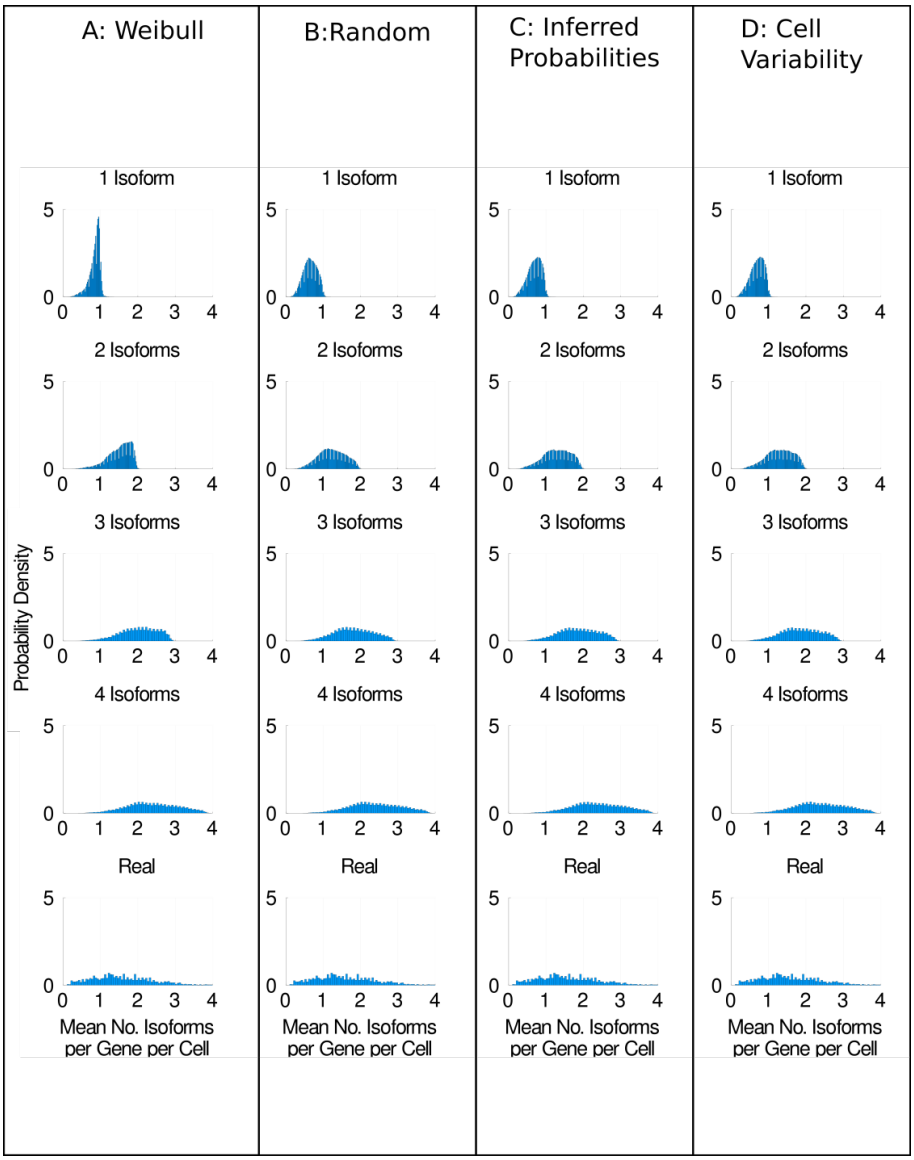


Figure 9.3: Different models of isoform choice alter our ability to detect isoforms. **A** Distributions of the mean number of isoforms detected per gene per cell for H9 hESCs sequenced at approximately 4 million reads per cell using the Weibull model of isoform choice (Bacher et al., 2017; Hu et al., 2017). **B** shows the same distributions when the random model is used. **C** shows the distributions when the inferred probabilities model is used. **D** shows the distributions when the cell variability model is used. See the main text for a detailed description of each model.

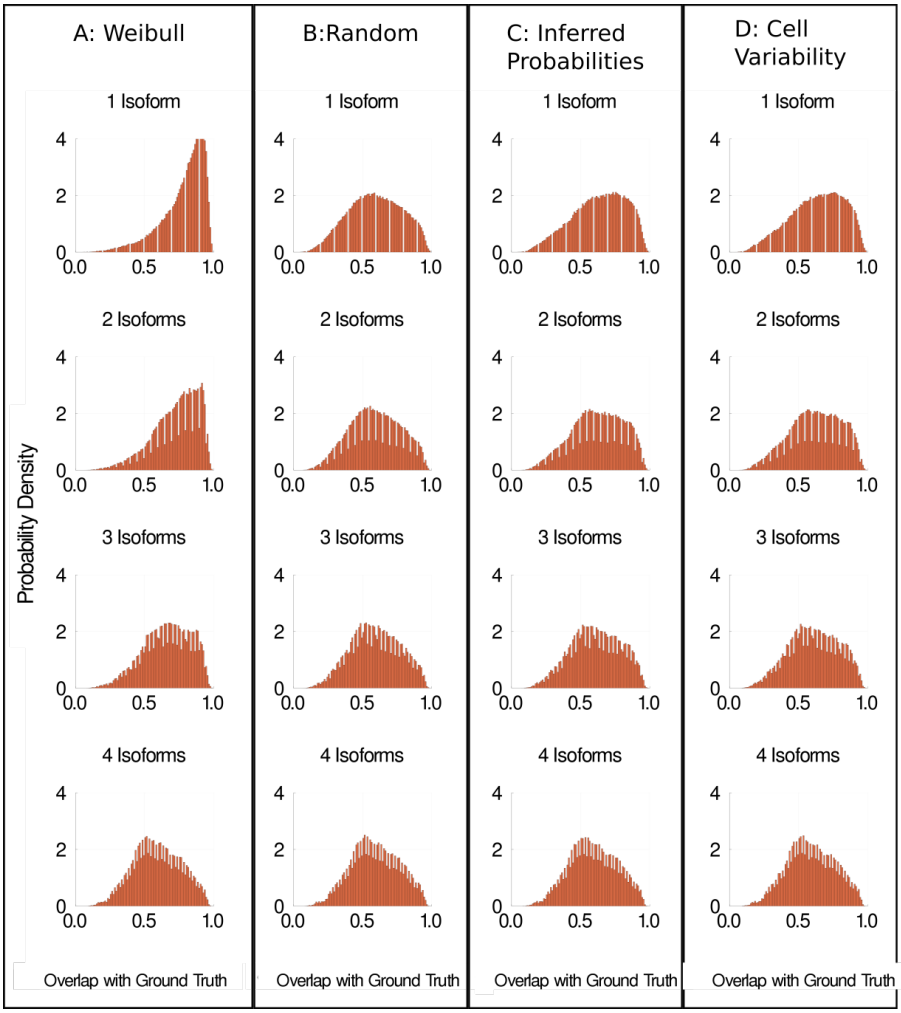


Figure 9.4: Different models of isoform choice alter our ability to detect isoforms. **a** Distributions of overlap fraction with the ground truth for H9 hESCs sequenced at approximately 4 million reads per cell using the Weibull model of isoform choice (Bacher et al., 2017; Hu et al., 2017). **b** shows the same distributions when the random model is used. **c** shows the distributions when the inferred probabilities model is used. **d** shows the distributions when the cell variability model is used. See the main text for a detailed description of each model.

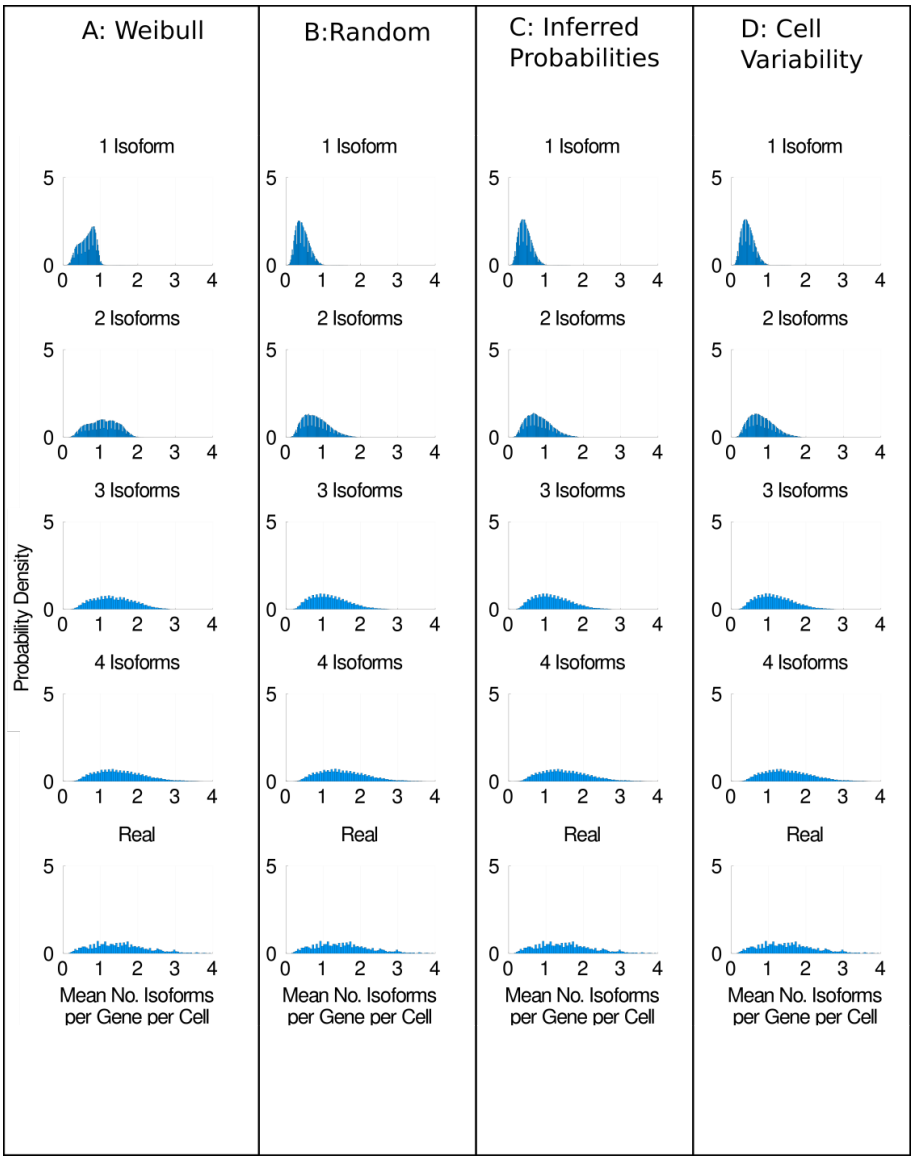


Figure 9.5: Different models of isoform choice alter our ability to detect isoforms. **A** Distributions of the mean number of isoforms detected per gene per cell for H9 hESCs sequenced at approximately 1 million reads per cell using the Weibull model of isoform choice (Bacher et al., 2017; Hu et al., 2017). **B** shows the same distributions when the random model is used. **C** shows the distributions when the inferred probabilities model is used. **D** shows the distributions when the cell variability model is used. See the main text for a detailed description of each model.

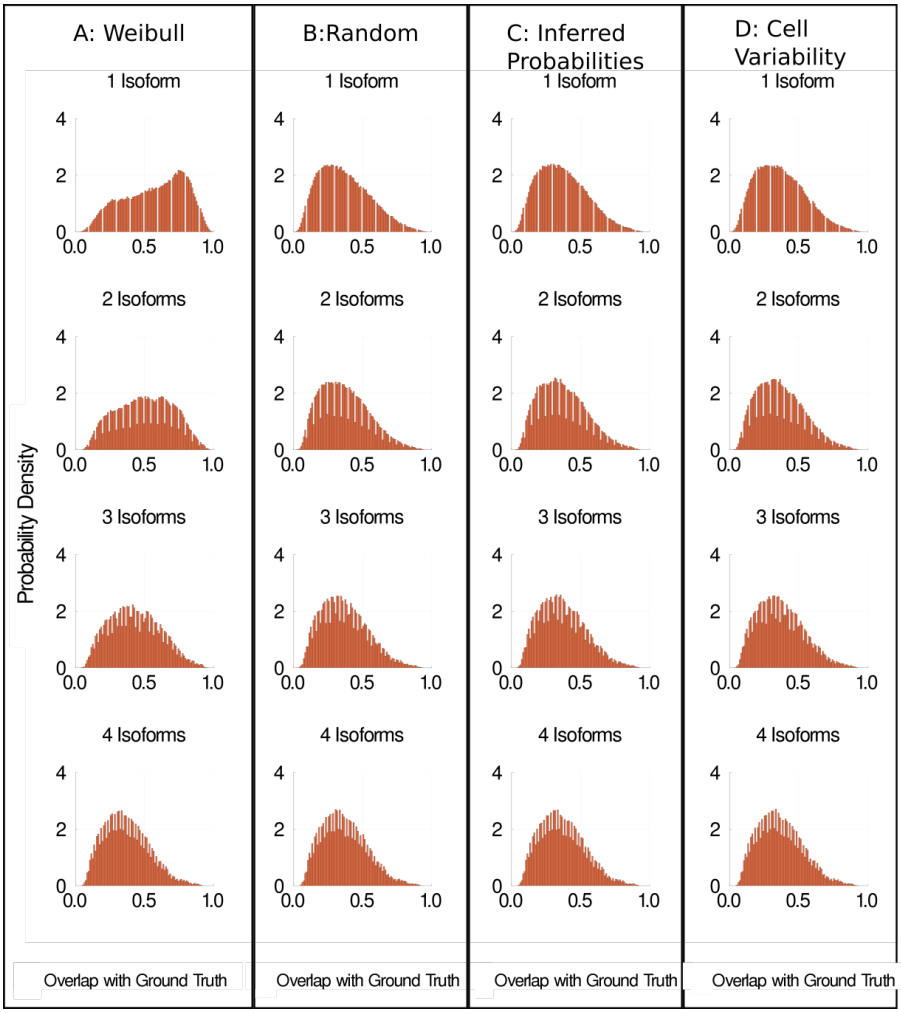


Figure 9.6: Different models of isoform choice alter our ability to detect isoforms. **a** Distributions of overlap fraction with the ground truth for H9 hESCs sequenced at approximately 1 million reads per cell using the Weibull model of isoform choice (Bacher et al., 2017; Hu et al., 2017). **b** shows the same distributions when the random model is used. **c** shows the distributions when the inferred probabilities model is used. **d** shows the distributions when the cell variability model is used. See the main text for a detailed description of each model.

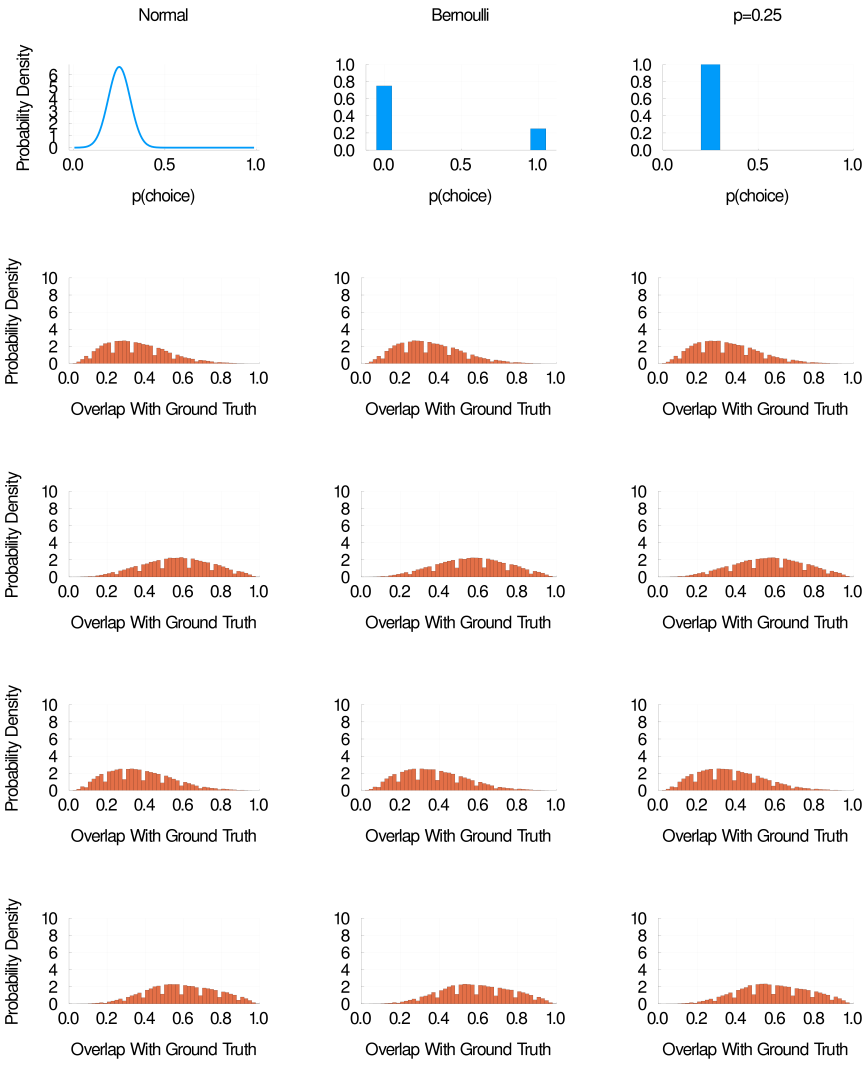


Figure 9.7: Some models of isoform choice are more plausible than others. I model the probability of picking any given isoform as a Normal distribution, a Bernoulli distribution and a constant probability, all with the same mean (0.25) (top row of graphs). In the following rows, I show the distributions of the overlap fraction when each model of isoform choice is used. The second row is H1 hESCs sequenced at 1 million reads per cell, the third row is H1 hESCs sequenced at 4 million reads, the fourth row is H9 hESCs sequenced at 1 million reads, the fifth row is H9 hESCs sequenced at 4 million reads.

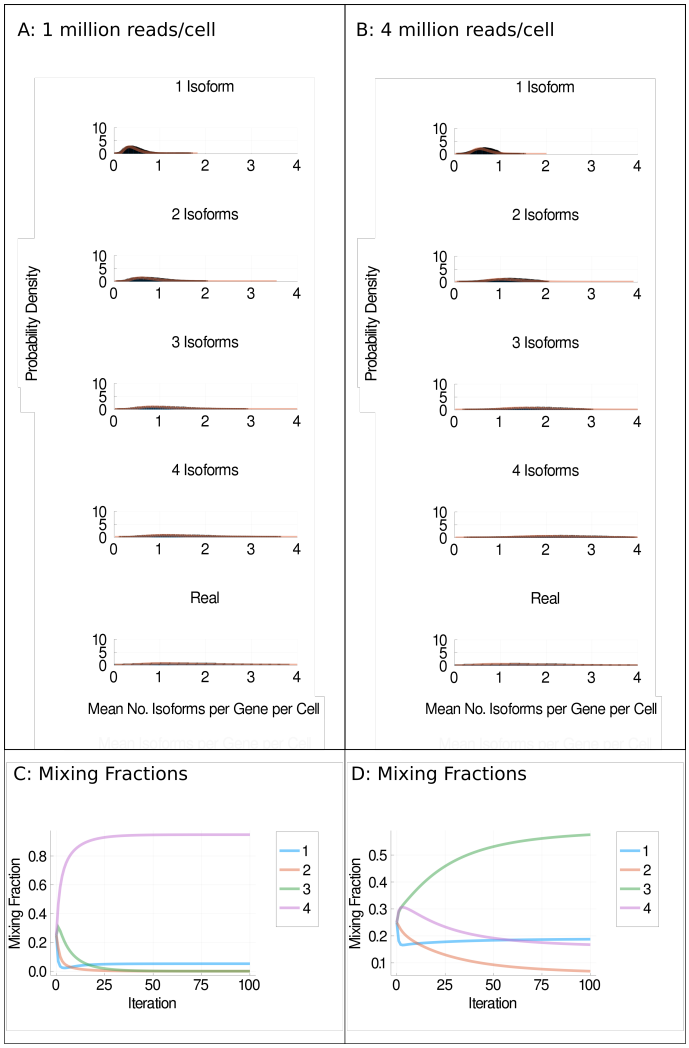


Figure 9.8: Mixture models. **A** and **B** Distributions of detected isoforms per gene per cell (blue) and log normal fitted distributions (orange) for H1 cells sequenced at 1 million reads per cell (**A**) or 4 million reads per cell (**B**) under the random model (Bacher et al., 2017). **C** and **D** Mixing fractions vs iterations of expectation maximisation for 1 million reads per cell (**C**) and 4 million reads per cell (**D**). Each coloured line represents the distributions for one, two, three or four isoforms being simulated as expressed per gene per cell.

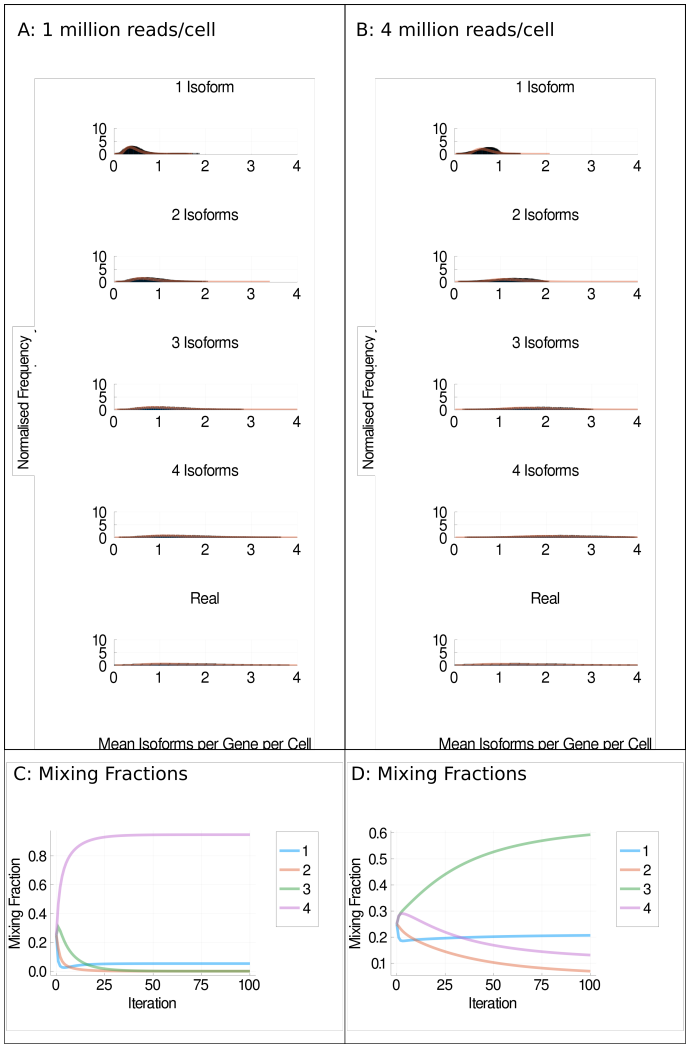


Figure 9.9: Mixture models. **A** and **B** Distributions of detected isoforms per gene per cell (blue) and log normal fitted distributions (orange) for H1 cells sequenced at 1 million reads per cell (**A**) or 4 million reads per cell (**B**) under the inferred model (Bacher et al., 2017). **C** and **D** Mixing fractions vs iterations of expectation maximisation for 1 million reads per cell (**C**) and 4 million reads per cell (**D**). Each coloured line represents the distributions for one, two, three or four isoforms being simulated as expressed per gene per cell.

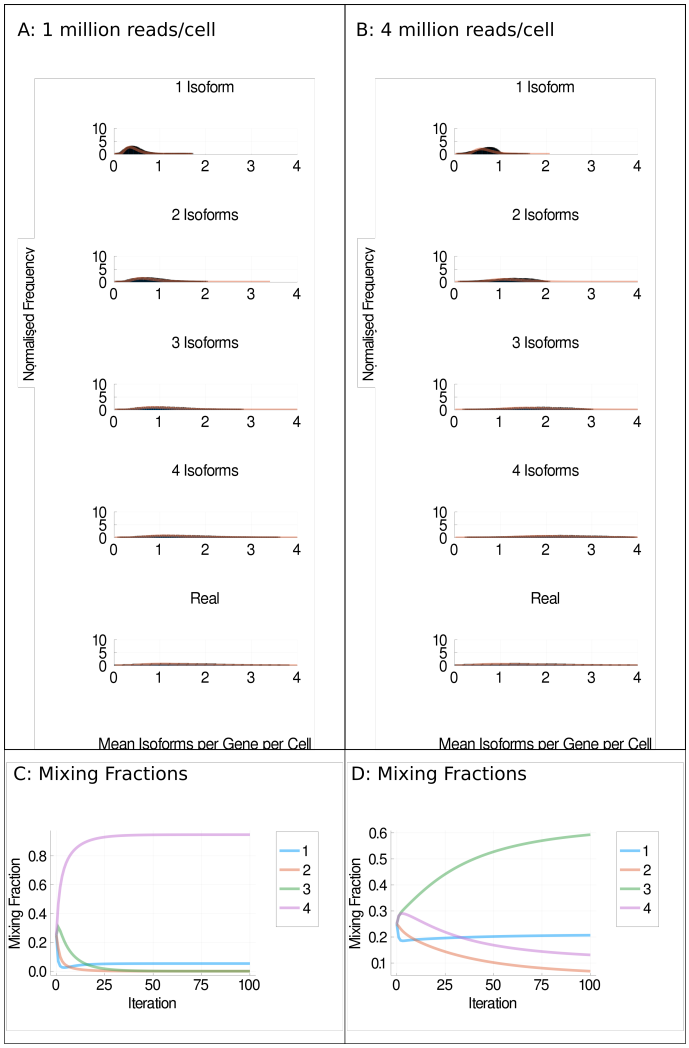


Figure 9.10: Mixture models. **A** and **B** Distributions of detected isoforms per gene per cell (blue) and log normal fitted distributions (orange) for H1 cells sequenced at 1 million reads per cell (**A**) or 4 million reads per cell (**B**) under the cell variable model (Bacher et al., 2017; Velten et al., 2015). **C** and **D** Mixing fractions vs iterations of expectation maximisation for 1 million reads per cell (**C**) and 4 million reads per cell (**D**). Each coloured line represents the distributions for one, two, three or four isoforms being simulated as expressed per gene per cell.

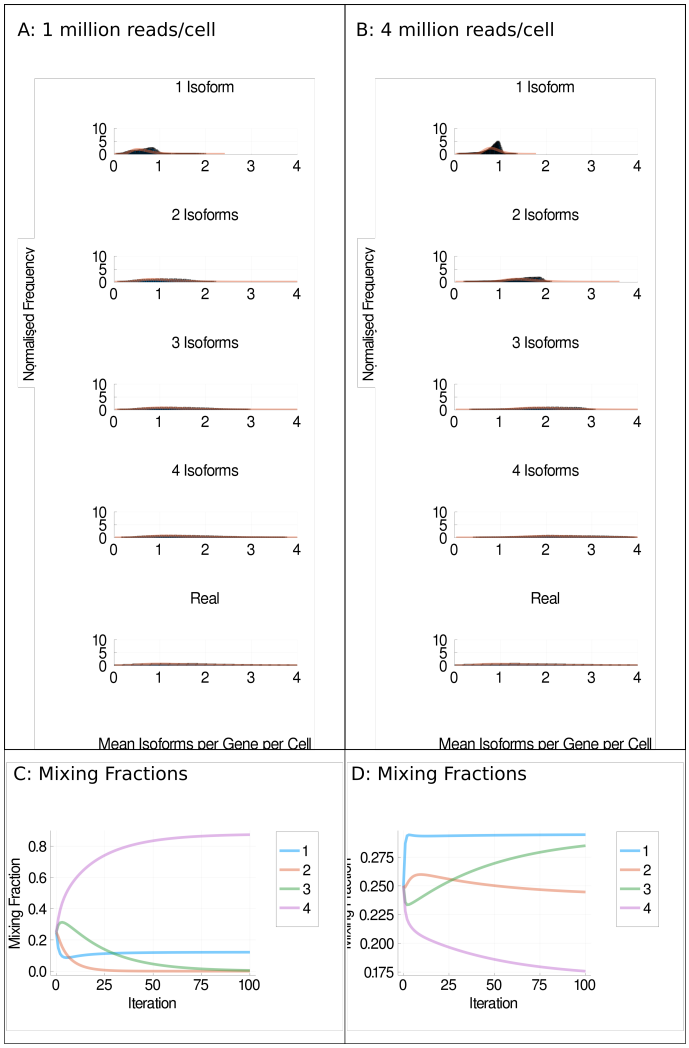


Figure 9.11: Mixture models. **A** and **B** Distributions of detected isoforms per gene per cell (blue) and log normal fitted distributions (orange) for H9 cells sequenced at 1 million reads per cell (**A**) or 4 million reads per cell (**B**) under the Weibull model (Bacher et al., 2017; Hu et al., 2017). **C** and **D** Mixing fractions vs iterations of expectation maximisation for 1 million reads per cell (**C**) and 4 million reads per cell (**D**). Each coloured line represents the distributions for one, two, three or four isoforms being simulated as expressed per gene per cell.

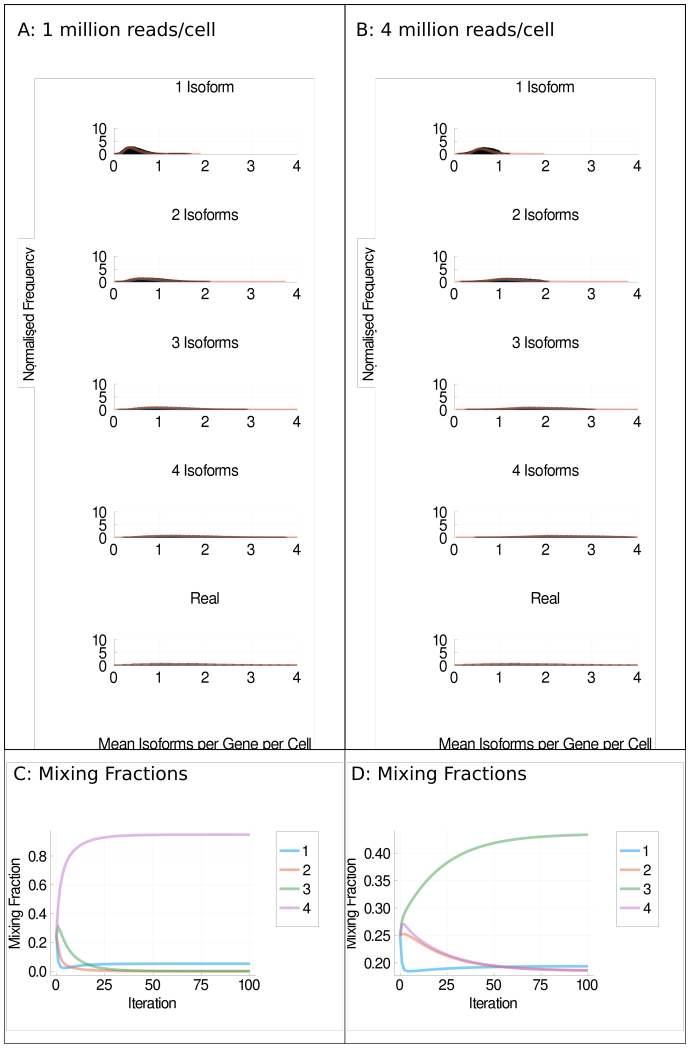


Figure 9.12: Mixture models. **A** and **B** Distributions of detected isoforms per gene per cell (blue) and log normal fitted distributions (orange) for H9 cells sequenced at 1 million reads per cell (**A**) or 4 million reads per cell (**B**) under the random model (Bacher et al., 2017). **C** and **D** Mixing fractions vs iterations of expectation maximisation for 1 million reads per cell (**C**) and 4 million reads per cell (**D**). Each coloured line represents the distributions for one, two, three or four isoforms being simulated as expressed per gene per cell.

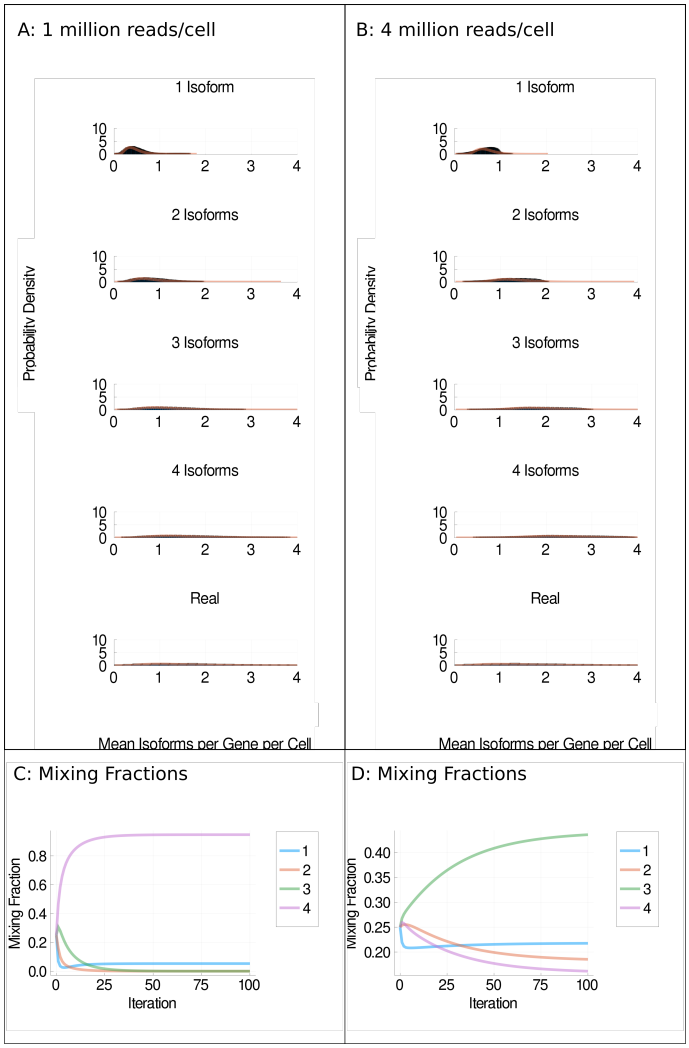


Figure 9.13: Mixture models. **A** and **B** Distributions of detected isoforms per gene per cell (blue) and log normal fitted distributions (orange) for H9 cells sequenced at 1 million reads per cell (**A**) or 4 million reads per cell (**B**) under the inferred model (Bacher et al., 2017). **C** and **D** Mixing fractions vs iterations of expectation maximisation for 1 million reads per cell (**C**) and 4 million reads per cell (**D**). Each coloured line represents the distributions for one, two, three or four isoforms being simulated as expressed per gene per cell.

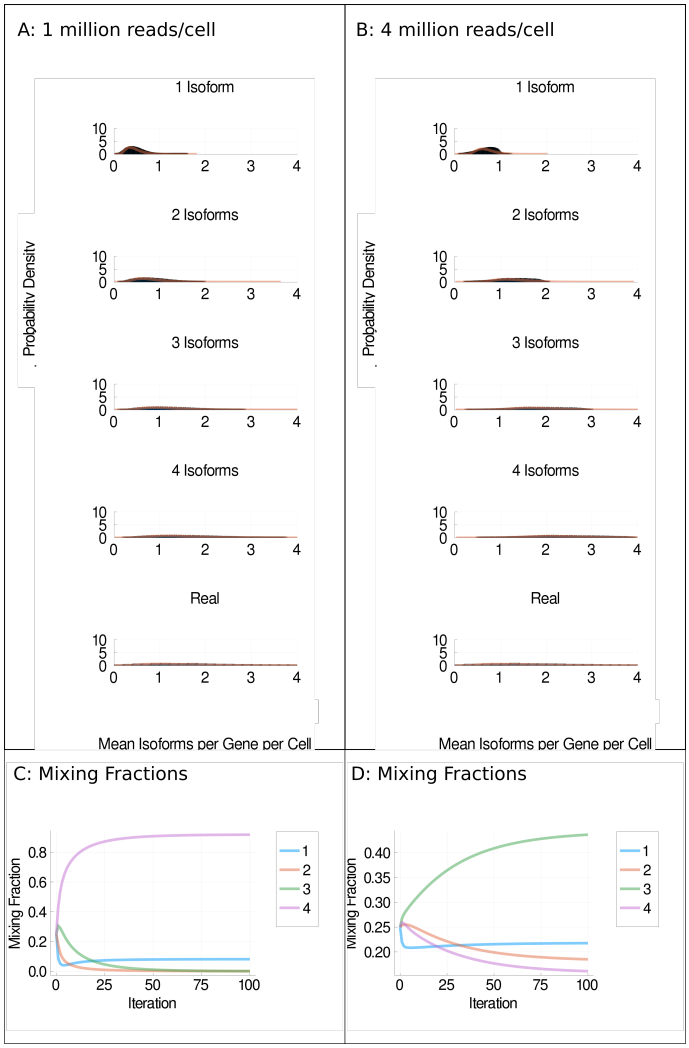


Figure 9.14: Mixture models. **A** and **B** Distributions of detected isoforms per gene per cell (blue) and log normal fitted distributions (orange) for H9 cells sequenced at 1 million reads per cell (**A**) or 4 million reads per cell (**B**) under the cell variable model (Bacher et al., 2017; Velten et al., 2015). **C** and **D** Mixing fractions vs iterations of expectation maximisation for 1 million reads per cell (**C**) and 4 million reads per cell (**D**). Each coloured line represents the distributions for one, two, three or four isoforms being simulated as expressed per gene per cell.

## 9.1 Supplementary Tables

| No. Isoforms Simulated | p-Value |
|---|---|
| 1 | 0.0 |
| 2 | 0.0 |
| 3 | 0.0 |
| 4 | 0.999999 |

Table 9.1: Results of K-sample Anderson–Darling test, which tests whether multiple collections come from the same population. The test was applied to each row of graphs in Figure 4.13, in other words testing whether the distributions generated by different isoform choice models are significantly different.

| No. Isoforms Simulated | p-Value |
|---|---|
| 1 | 0.835737 |
| 2 | 0.997938 |
| 3 | 0.998721 |
| 4 | 0.99074 |

Table 9.2: Results of K-sample Anderson–Darling test, which tests whether multiple collections come from the same population. The test was applied to the simulation results generated using the Inferred Probabilities vs the Cell Variable models of isoform choice in Figure 4.13 to test whether the distributions generated by different isoform choice models significantly differ.

| No. Isoforms Simulated | p-Value |
|---|---|
| 1 | 0.0 |
| 2 | 0.0 |
| 3 | 0.0 |
| 4 | 1.0 |

Table 9.3: Results of K-sample Anderson–Darling test, which tests whether multiple collections come from the same population. The test was applied to each row of graphs in Figure 4.12, in other words testing whether the distributions generated by different isoform choice models are significantly different.

| No. Isoforms Simulated | p-Value |
|---|---|
| 1 | 0.639939 |
| 2 | 0.959654 |
| 3 | 0.995236 |
| 4 | 0.999814 |

Table 9.4: Results of K-sample Anderson–Darling test, which tests whether multiple collections come from the same population. The test was applied to the simulation results generated using the Inferred Probabilities vs the Cell Variable models of isoform choice in Figure 4.12 to test whether the distributions generated by different isoform choice models significantly differ.

| No. Isoforms Simulated | p-Value |
|---|---|
| 1 | 0.0 |
| 2 | 0.0 |
| 3 | 0.0 |
| 4 | 0.999999 |

Table 9.5: Results of K-sample Anderson–Darling test, which tests whether multiple collections come from the same population. The test was applied to each row of graphs in Supplementary Figure 9.3, in other words testing whether the distributions generated by different isoform choice models are significantly different.

| No. Isoforms Simulated | p-Value |
|---|---|
| 1 | 0.98348 |
| 2 | 0.95075 |
| 3 | 0.999405 |
| 4 | 0.995485 |

Table 9.6: Results of K-sample Anderson–Darling test, which tests whether multiple collections come from the same population. The test was applied to the simulation results generated using the Inferred Probabilities vs the Cell Variable models of isoform choice in Figure 9.3 to test whether the distributions generated by different isoform choice models significantly differ.

| No. Isoforms Simulated | p-Value |
|---|---|
| 1 | 0.0 |
| 2 | 0.0 |
| 3 | 0.0 |
| 4 | 1.0 |

Table 9.7: Results of K-sample Anderson–Darling test, which tests whether multiple collections come from the same population. The test was applied to each row of graphs in Supplementary Figure 9.5, in other words testing whether the distributions generated by different isoform choice models are significantly different.

| No. Isoforms Simulated | p-Value |
|---|---|
| 1 | 0.932755 |
| 2 | 0.969666 |
| 3 | 0.999973 |
| 4 | 0.999753 |

Table 9.8: Results of K-sample Anderson–Darling test, which tests whether multiple collections come from the same population. The test was applied to the simulation results generated using the Inferred Probabilities vs the Cell Variable models of isoform choice in Supplementary Figure 9.5 to test whether the distributions generated by different isoform choice models significantly differ.

| Data source | p-Value |
|---|---|
| H1 1 million reads | 0.99808 |
| H1 4 million reads | 0.981612 |
| H9 1 million reads | 0.989299 |
| H9 4 million reads | 0.997866 |

Table 9.9: Results of K-sample Anderson–Darling test, which tests whether multiple collections come from the same population. The test was applied to the simulation results generated using the Normal, Bernoulli and p=0.25 models of isoform choice to test whether the distributions generated by different isoform choice models significantly differ.